NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

EVALUATING ATLANTIC TROPICAL CYCLONE TRACK ERROR DISTRIBUTIONS BASED ON FORECAST CONFIDENCE

by

Matthew D. Hauke

June 2006

Thesis Advisor:	Patrick A. Harr
Second Reader:	Russell L. Elsberry

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2006	Master's Thesis

4. TITLE AND SUBTITLE Evaluating Atlantic Tropical Cyclone Track Error Distributions Based on Forecast Confidence	5. FUNDING NUMBERS
6. AUTHOR(S) Matthew D. Hauke	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

A new Tropical Cyclone (TC) surface wind speed probability product from the National Hurricane Center (NHC) takes into account uncertainty in track, maximum wind speed, and wind radii. A Monte Carlo (MC) model is used that draws from probability distributions based on historic track errors. In this thesis, distributions of forecast track errors conditioned on forecast confidence are examined to determine if significant differences exist in distribution characteristics. Two predictors are used to define forecast confidence: the Goerss Predicted Consensus Error (GPCE) and the Global Forecast System (GFS) ensemble spread. The distributions of total-, along-, and cross-track errors from NHC official forecasts are defined for low, average, and high forecast confidence. Also, distributions of the GFS ensemble mean total-track errors are defined based on similar confidence levels. Standard hypothesis testing methods are used to examine distribution characteristics. Using the GPCE values, significant differences in nearly all track error distributions existed for each level of forecast confidence. The GFS ensemble spread did not provide a basis for statistically different distributions. These results suggest that the NHC probability model would likely be improved if the MC model would draw from distributions of track errors based on the GPCE measures of forecast confidence

14. SUBJECT TERMS Tropical Cyclone, Track Errors, CONU, Consensus, GPCE, Forecast Confidence, National Hurricane Center, Monte Carlo Model, Probabilistic Forecast, Wind Speed Probability	15. NUMBER OF PAGES 105
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

EVALUATING ATLANTIC TROPICAL CYCLONE TRACK ERROR DISTRIBUTIONS BASED ON FORECAST CONFIDENCE

Matthew D. Hauke
Captain, United States Air Force
B.S., University of Wisconsin, 1997

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN METEOROLOGY

from the

NAVAL POSTGRADUATE SCHOOL
June 2006

Author: Matthew D. Hauke

Approved by: Patrick A. Harr
Thesis Advisor

Russell L. Elsberry
Second Reader

Philip A. Durkee
Chairman, Department of Meteorology

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

A new Tropical Cyclone (TC) surface wind speed probability product from the National Hurricane Center (NHC) takes into account uncertainty in track, maximum wind speed, and wind radii. A Monte Carlo (MC) model is used that draws from probability distributions based on historic track errors. In this thesis, distributions of forecast track errors conditioned on forecast confidence are examined to determine if significant differences exist in distribution characteristics. Two predictors are used to define forecast confidence: the Goerss Predicted Consensus Error (GPCE) and the Global Forecast System (GFS) ensemble spread. The distributions of total-, along-, and cross-track errors from NHC official forecasts are defined for low, average, and high forecast confidence. Also, distributions of the GFS ensemble mean total-track errors are defined based on similar confidence levels. Standard hypothesis testing methods are used to examine distribution characteristics. Using the GPCE values, significant differences in nearly all track error distributions existed for each level of forecast confidence. The GFS ensemble spread did not provide a basis for statistically different distributions. These results suggest that the NHC probability model would likely be improved if the MC model would draw from distributions of track errors based on the GPCE measures of forecast confidence.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACRONYMS

ATCF	Automated Tropical Cyclone Forecast system
AVNI	"Aviation" run of the NCEP GFS - Interpolated
CONU	At least two of GFDI, AVNI, NGPI, UKMI, and GFNI model ensemble average
GFDI	GFDL - Interpolated
GFDL	Geophysical Fluid Dynamics Laboratory model
GFNI	Navy version of GFDL - Interpolated
GFS	Global Forecasting System
GPCE	Goerss Predicted Consensus Error
GUNA	GFDI-UKMI-NGPI-AVNI model ensemble average
GUNS	GFDL-UKMET-NOGAPS model ensemble average
HURCON	Hurricane Condition
MC	Monte Carlo Model
NASA	National Aeronautics and Space Administration
NCEP	National Centers for Environmental Prediction
NHC	National Hurricane Center
NGPI	NOGAPS - Interpolated
NOAA	National Oceanic and Atmospheric Administration
NOGAPS	Naval Operational Global Atmospheric Prediction System
NWS	National Weather Service
OFCL	Official NHC Forecast
TC	Tropical Cyclone
UKMI	United Kingdom Meteorological Office model – Interpolated

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank several people for their contributions and guidance for this thesis. First, I would like to thank my advisor, Professor Patrick A. Harr. Without his guidance and programming ability this thesis would have been difficult to accomplish. Next, I would like to thank Mr. William Roeder of the 45th Weather Squadron, Patrick AFB, Florida whose ideas and guidance were instrumental in starting this process. Thanks is also due to Dr. Mark DeMeria from NESDIS for his guidance, Dr. James Goerss and Mr. Buck Sampson from NRL for their guidance and data, and Mr. James Franklin of NHC for the along- and cross- track errors. Finally, I would like to thank Professor Russell L. Elsberry for his guidance and expertise.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. MOTIVATION AND OBJECTIVE

In both the civilian and military world, weather phenomena can be classified by degrees of impact. An example of a low degree of impact can be a crosswind (component of the wind blowing perpendicular to the landing surface) of 20 knots. This may ground certain types of sensitive aircraft such as a U-2, but a Boeing 747 or a military C-17 can continue to operate. This weather phenomenon can occur often, but its impact is minimal. However, a weather phenomenon such as a Tropical Cyclone (TC) can have dangerous winds, flood-causing heavy rain, and destructive storm surge over large areas so the degree of impact will be high. An approaching hurricane can curtail the civilian economy, military operations, and all routine day-to-day living along several hundred miles of coastline and inland areas. The duration of the impact can be several days to several months depending on the damage. This weather phenomenon rarely occurs but its impacts can be total devastation.

The 2004 and 2005 Atlantic hurricane seasons demonstrated all too well the impacts of hurricanes on both civilian and military centers. For example, in 2005, Hurricane Katrina destroyed several hundred miles of coastline including the city of New Orleans and Keesler Air Force Base (AFB). Destruction in New Orleans cost billions of dollars and caused more then 1,000 deaths. It will take many years to restore the city to its pre-Katrina status. Keesler AFB experienced a large storm surge that caused heavy damage to almost every structure. As a training hub for the Air Force, Keesler AFB trains thousands of Airmen every year, especially new personnel receiving their initial career training. With the destruction caused from Hurricane Katrina, the mission was forced to stop for a few months, which caused a huge impact on the Air Force in general.

The best information available needs to be supplied to the civilian and military worlds to protect lives and allow preparations that will minimize the damage. The goal of this thesis is to improve this information.

A 2003 Joint Hurricane Testbed project funded Dr. Mark DeMaria, Dr. John Knaff, and colleagues to transfer to operations a new probabilistic product. This product

uses a new statistical model to determine the probabilities that certain wind speed thresholds will be exceeded at certain points. This project represented a radical departure from the old way of probabilistic TC forecasting.

The goal of this thesis is to investigate whether improvements to this new probabilistic model could be made by introducing different distributions of track forecast errors that the model utilizes to calculate probabilities. If it is possible to use different distributions for different situations, the probabilistic output may be more representative. Such improved guidance could lead to a reduction in the massive costs of overly cautious evacuations when track forecast confidence is high, or even save lives by expanding the necessary evacuation zone when forecast confidence is low. For this thesis, the following hypothesis will be investigated:

> Track forecast error distributions may be altered by considering forecast uncertainty (i.e., difficulty), which is defined by the variations among the track forecast aids, and introduction of these track forecast error distributions by forecast interval will positively influence the strike probability distributions along the track.

B. 2005 ATLANTIC HURRICANE SEASON

Perhaps no other Atlantic hurricane season in history stressed the importance of timely and accurate forecasts than the 2005 season. With 28 tropical storms and an additional three depressions, the 2005 Atlantic hurricane season was the most active on record (see Figure 1). The season broke the previous record of 21 named storms set in 1933. In addition, 15 of the 28 tropical storms intensified to hurricane strength, which broke the record of 12 set in 1969. To further illustrate the severity of the 2005 Atlantic hurricane season, seven of the 13 hurricanes became major hurricanes (Category 3 or higher), four of which made landfall in the United States (Dennis, Katrina, Rita, and Wilma). Also including in the 13 major hurricanes were four Category 5 storms (Emily, Katrina, Rita, and Wilma), which was one more than the previous record of three Category 5 storms (NHC 2006).

The damages during the 2005 Atlantic hurricane season were estimated to be $150 billion. This total broke the prior damage records set in 1992 and 2004, which each

had $50 billion in damage when adjusted to 2005 dollars. It was also the deadliest hurricane season for the U.S. since 1928 (Wilson 2006).

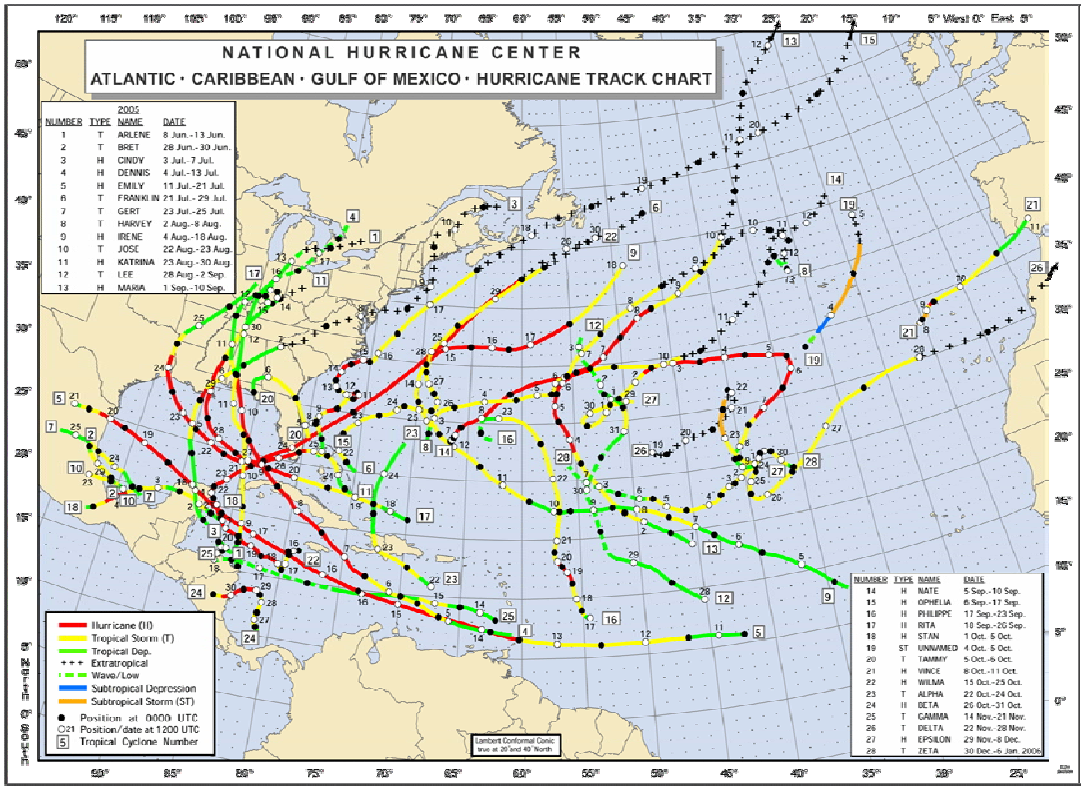

Figure 1. 2005 Atlantic hurricane season track map (from NHC, http://www.nhc.noaa.gov/tracks/2005atl.gif).

Some other notable records broken during the 2005 Atlantic hurricane season include the strongest hurricane (Wilma 882 mb), three of the six strongest on record (Wilma 882 mb, Rita 897 mb, and Katrina 902 mb), top sustained winds (Emily at 160 mph), and the longest-lasting hurricane for the month of December (Epsilon) (Wilson 2006).

The 2005 hurricane season reinforced the importance of TC forecasting, preparation, and evacuation in advance of landfall. The devastation caused by Hurricane Katrina captured the public's attention in a way that no hurricane has been able to do since the storm that unexpectedly hit Galveston, Texas in 1900 and killed over 6000 people. Unfortunately, history has shown that sometimes it takes a catastrophe to initiate dramatic progress for the better. The hope is that the lessons learned from Katrina will save more lives in the future than were lost in this disaster.

C. CIVILIAN IMPACTS AND PROCEDURES

In advance of a hurricane, the public has little control of the eventual outcome. When required, the best they can do is to shore up property, collect personal belongings, evacuate the area, and then hope for the best.

Emergency managers use an estimate of economic impact of about $1 million per mile of coastline for evacuation. This number is highly subjective and studies have shown this number to be dependent on several factors, including population, economic class, and storm intensity. Whitehead (2000) found it cost between $1 million to $50 million to evacuate the North Carolina (NC) coastline depending on storm intensity and cost of evacuation plans. Since there are more than 50 miles of NC coastline (Figure 2), even the worse-case scenarios have the cost of evacuation less than $1 million per mile. Either way, evacuation is a costly proposition in preparation, transportation, and lodging, and the shutting down of the local economy for an unknown number of days. The cost/risk analysis is different for every evacuation zone, which makes the evacuation declaration a hard decision to make.

Given the human lives and economic factors involved, the NHC takes its watch/warning advisories very seriously. A missed forecast may lead to a whole population center being at risk, and a false alarm can cost a lot of money and decrease public confidence in their forecasts. Hurricane Katrina was a perfect example of how public confidence can change over the years. For years, the doomsday scenario had been possible for New Orleans each time a hurricane approached that section of coast. In nearly every case, the storm would either miss or not be strong enough to cause significant damage, which may have caused some of the public to ignore the NHC advisories in advance of Katrina. The hope was that it would veer and miss just as had occurred plenty of times in the past. Three days before landfall, a lot of people were still lackadaisical about the threat, even though the NHC had high confidence in its forecast that New Orleans was likely to experience a direct hit. The end result was that many of people were caught off guard when the impact became imminent. By the time landfall near New Orleans was imminent, there was not enough time to evacuate for those who stayed behind.

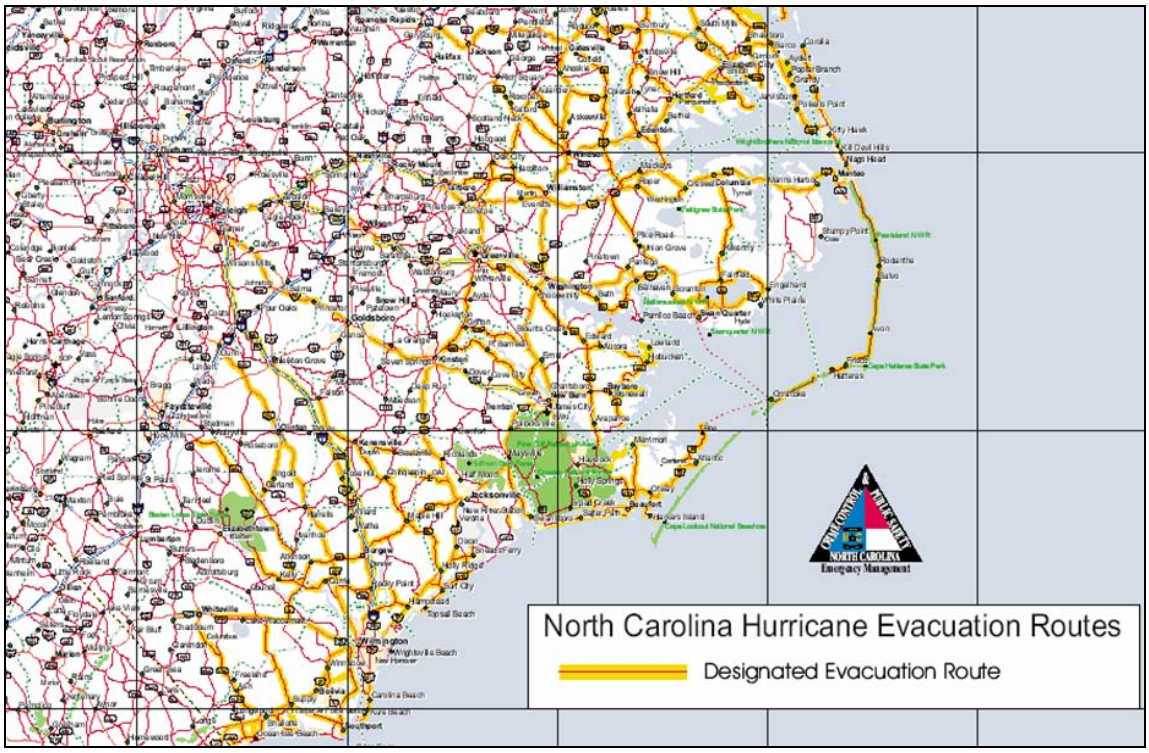

Figure 2. North Carolina Hurricane Evacuation Routes map is an example of the complex planning involved for evacuating coastlines (from NOAA's Hurricane Evacuation Zone Maps archive, http://www.dem.dcc.state.nc.us/hurricane/HurricaneEvacuationRoutes.pdf).

The NHC is continually looking for ways to decrease the distance along the coast for their watch/warnings as their guidance products become more accurate. As previously mentioned, too many false alarms will lower the public's confidence in the forecasts, especially because of the high costs and inconvenience involved with evacuation. The goal of this thesis is to contribute to the solution of these problems by improving hurricane probabilistic predictions. A decrease in false alarms directly saves the public money and indirectly saves lives by increasing the public's confidence in the forecasts.

D. MILITARY IMPACTS AND PROCEDURES

1. TC Impacts on the Military

Every military installation along the Eastern Seaboard, Gulf of Mexico, and in the Atlantic Basin has some type of hurricane preparedness plan. Each installation is unique in its plans, weather thresholds, and reaction to approaching TCs.

Air Force Bases, Army Air Fields, and Navy and Marine Airfields all have evacuation plans to fly aircraft out of the storm's destructive path. In addition, military ships need to steer clear of the storm or sortie from a port, military installations need to be prepared, and personnel need time to prepare their property and families to evacuate. All of this preparation can cost money and a false alarm can cripple a budget, but a missed forecast can be much more expensive and devastating. So commanders want the best information available to make their decision.

2. TC Impacts on the 45th Weather Squadron (WS), Patrick AFB, Florida

The 45 WS has perhaps the most sensitive mission in all of the U.S. military when it comes to hurricane preparedness. The NASA Space Shuttle and other space launch vehicles that are launched from Cape Canaveral Air Force Station (AFS) and Kennedy Space Center are unique in terms of importance, cost, ability to replace the resources, and the ability to reproduce the mission elsewhere (Figure 3). The 45 WS provides all of operational weather support to Cape Canaveral AFS and Kennedy Space Center. As home of the Space Shuttle along with a host of other rockets that often carry very expensive payloads, the 45 WS is required to provide highly specialized TC forecasts (Winters et al. 2006).

Because of the time needed to prepare the Space Shuttle for a possible hurricane strike, preparation starts days in advance. In the daily briefing with the Kennedy Space Center, the 45 WS reports any tropical activity in the Atlantic Basin. If it looks as though a possibility exists that a TC will impact operations, then several escalating steps are made as a strike becomes more imminent.

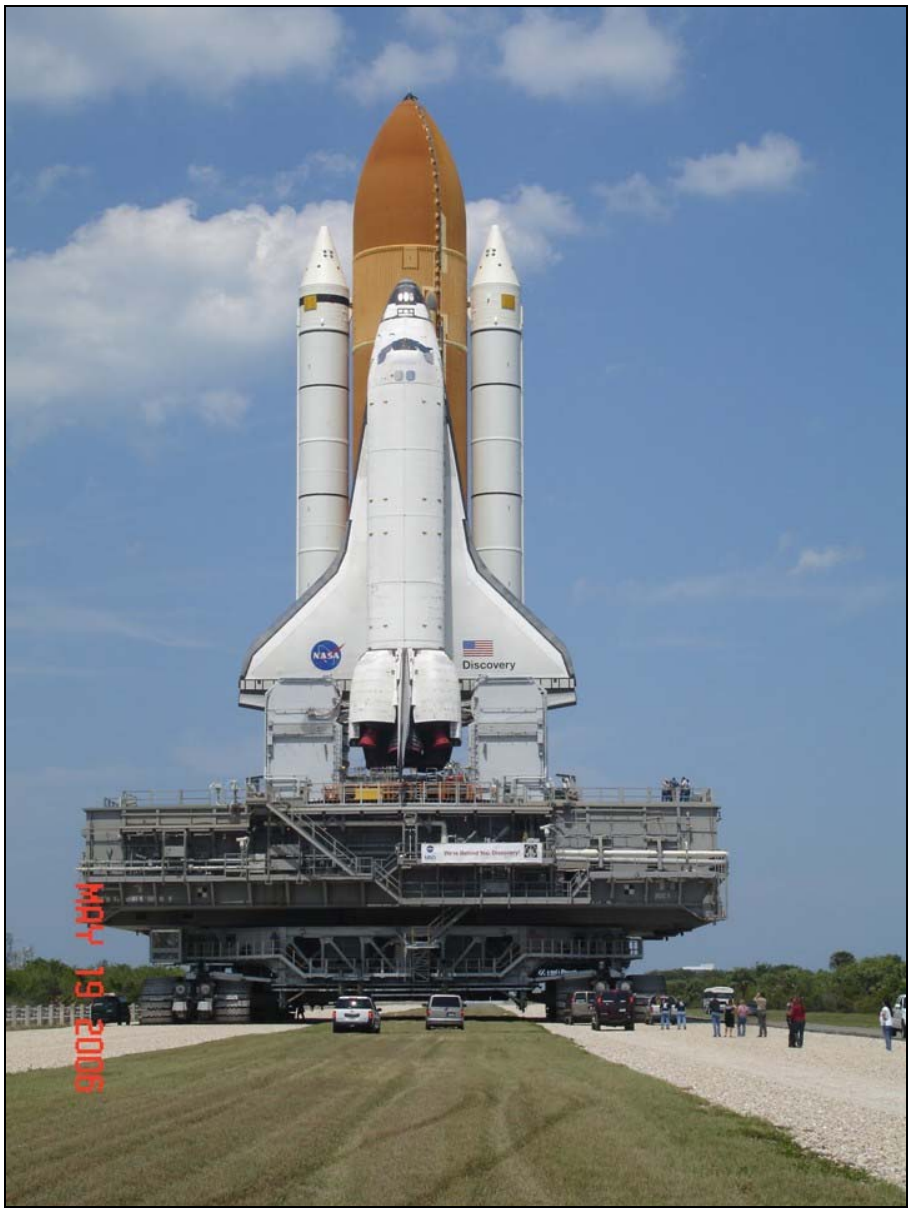

Figure 3. Space Shuttle rollback from the launch pad to the Vehicle Assembly Building (Provided by William Roeder of the 45 WS).

Interestingly, the 45 WS has communicated the threat of potentially hazardous TC weather to their customers in a probabilistic manner for years. Although some customers want a definite yes or no answer, the uncertainty that is inherent in forecasting the atmosphere makes this an impossible proposition. The NASA decision makers actually want a probabilistic forecast. Specifying probabilities allow the forecaster to communicate the uncertainty in the forecast to the customer so that they can use probabilistic decision making to minimize the costs, risks, and expected impacts.

The probabilities that the 45 WS provides have to be the best and most accurate available. A missed forecast can cause billions of dollars in damage to the Space Shuttle, other space launch vehicles, and payloads. On the other hand, a false alarm can be an expensive mistake too in costs to relocate the space launch vehicles to protective shelters, risks of damage during transport, and lost launch opportunities in the range schedule.

The Space Shuttle and the launch pad can safely ride out wind gusts of 70 kt. If wind gusts greater than 70 kt are anticipated, then the Space Shuttle can rollback to the Vehicle Assembly Building where it can safely sustain steady winds up to 113 kt. However, the rollback phase has its own weather restrictions, including less than 10% chance of lightning within 20 n mi, steady winds less than 40 kt and peak winds less than 60 kt. These weather conditions must also be forecast to allow a safe and successful rollback. While the rollback can be done in only 8 hours, preparations usually begin 39 h before the rollback, the decision is usually made 48 h before the start of the rollback. However, the formal decision process usually starts at least 72 h in advance of the rollback, with pre-planning beginning as early as 120 h in advance (Winters et al. 2006).

In addition to the space vehicle rollback decisions, the 45 WS also gives advice as to the Hurricane Condition (HURCON), preparation, aircraft relocation, and personnel decisions by the 45th Space Wing at Cape Canaveral AFS, Patrick AFB, and NASA at Kennedy Space Center. The HURCONs that are used by the military are based on the expected onset of 50-kt sustained winds from a tropical cyclone. A HURCON-IV, HURCON-III, HURCON-II, and HURCON-I means 50-kt winds are expected within 72 hours, 48 hours, 24 hours, and 12 hours, respectively.

When the new NHC probability products became available in 2005 to the public as an experimental product, the 45 WS began evaluating the product. This evaluation found the product to be extremely useful in that it provides a more objective method for producing probabilities that the winds will exceed their customer's thresholds. These objective products were deemed superior to the previous primarily subjective method used by the 45 WS.

Two main improvements were suggested by the 45 WS as a result of their evaluation: 1) performance verification of the probabilities for the various wind

thresholds for all the forecast intervals; and 2) investigating if the product could be improved by scaling the errors for each forecast interval for each TC by the forecast confidence as parameterized by the ensemble spread among the forecast models. If the ensemble spreads of TC location and intensity for today's forecast were only half of the historical spreads, then perhaps the statistical error model to be applied to today's forecast would be better built using only half of the historical average. The performance verification requested by the 45 WS would help them their customers on the proper decisions. For example, the 45 WS noticed that the probabilities at long forecast intervals were surprisingly low for locations that eventually experienced those winds, and such low probabilities could mislead their customers into delaying their preparations. A 3% probability of at least 50 kt steady wind at a location for a 120-h forecast may seem like a low risk, but it actually corresponds to a high risk. Thus, the 45 WS needs to know the verification rates of the NHC probability products as a function of forecast probability and forecast interval.

This thesis is primarily driven by the need of the 45 WS to continuously look for ways to improve their TC forecasts. A successful effort could possibly lead to TC forecast improvements for both the military and civilian worlds.

Background material regarding important concepts related to this thesis are provided in Chapter II. The methodology used for this study is described in Chapter III. The results of the study are presented in Chapter IV, and the conclusions and future recommendations are given in Chapter V.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. NATIONAL WEATHER SERVICE PROBABILISTIC TROPICAL CYCLONE FORECASTS

1. Prior NHC Methods to Convey Uncertainty in the Forecast

Starting in the early 1980s, the National Weather Service started issuing Watch/Warning graphical maps along with the TC advisories (Figure 4). This graphic display was established in part to help convey the uncertainty in TC forecasting to the public. The "cone," or white area, on both sides of the track is the area in which over the last 10 years the TC will verify 90% of the time given the average track forecast errors. At each 12-h forecast interval, a circle is drawn with a radius of the average 10-year track forecast error, and then the circles are connected to form a cone. In other words, this model assumes that the same average track forecast error at each forecast interval applies to all tropical cyclones. While this conveys uncertainty in the forecast track of the storm, it does not include uncertainty in the forecasts of intensity and radii of wind speed thresholds. In addition, the public tends to fixate on the forecast track of the TC, or the center "black line," while tending to ignore the cone on each side. Some people on the periphery of the cone have tended to delay evacuation until an actual shift in storm motion takes place, perhaps because they did not understand the actual probabilities of a strike. Thus, people on the periphery of the cone have been caught off guard several times in the past when the TC deviates from the forecast track. A great example of this situation is Hurricane Charley in 2004, which will be discussed in the next section.

The NHC also issues a strike probability forecast in text format (Figure 5). The probabilities are determined by the percentage of times a TC within the given time frame will pass within 75 nautical miles (n mi) to the right or 50 n mi to the left of a point relative to the direction of cyclone motion. These probabilities are also conditioned on historical tracks and do not consider different intensities and critical radii differences among storms. While this product may provide a good indicator of probability and uncertainty, the strike probability forecast only goes out to 72 h and is not a user-friendly product.

Because the strike probability forecast text products are not user-friendly and the Watch/Warning graphical product does not convey uncertainty in intensity and radii, a totally new method of probabilistic TC forecasting was created.

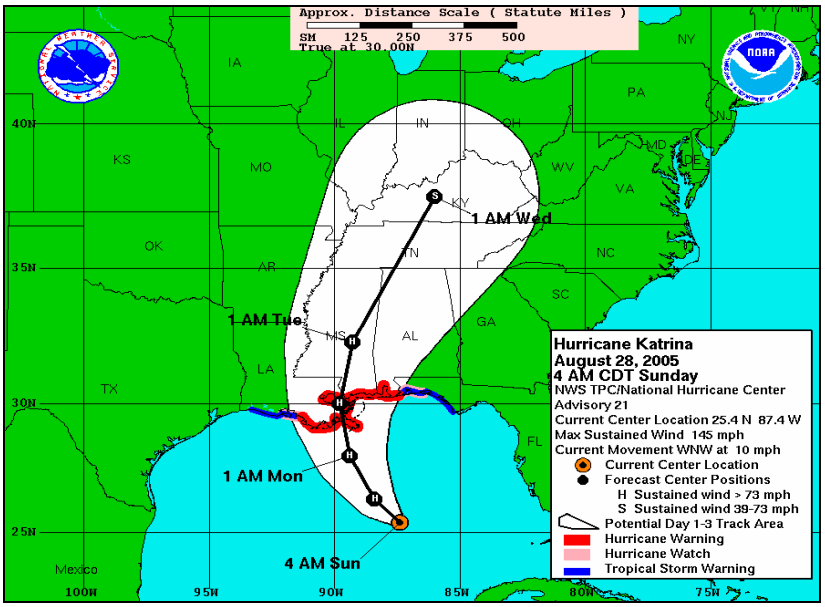

Figure 4. Watch/Warning 3-Day map for Katrina (from NHC, http://www.nhc.noaa.gov/archive/2005/KATRINA_graphics.shtml).

ZCZC MIASPFAT2 ALL
TTAA00 KNHC DDHHMM
HURRICANE KATRINA PROBABILITIES NUMBER 21
NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL
4 AM CDT SUN AUG 28 2005

PROBABILITIES FOR GUIDANCE IN HURRICANE PROTECTION
PLANNING BY GOVERNMENT AND DISASTER OFFICIALS

AT 4 AM CDT...0900Z...THE CENTER OF KATRINA WAS LOCATED NEAR
LATITUDE 25.4 NORTH...LONGITUDE 87.4 WEST

CHANCES OF CENTER OF THE HURRICANE PASSING WITHIN 65 NAUTICAL MILES
OF LISTED LOCATIONS THROUGH 1AM CDT WED AUG 31 2005

LOCATION	A	B	C	D	E	LOCATION	A	B	C	D	E
28.0N 89.4W	38	X	X	1	39	MOBILE AL	1	17	4	1	23
30.0N 89.8W	7	21	X	X	28	GULFPORT MS	3	21	2	X	26
32.3N 89.3W	X	7	14	1	22	BURAS LA	17	12	X	1	30
JACKSONVILLE FL	X	X	X	3	3	NEW ORLEANS LA	7	21	1	X	29
SAVANNAH GA	X	X	X	4	4	NEW IBERIA LA	3	20	1	X	24
CHARLESTON SC	X	X	X	3	3	PORT ARTHUR TX	X	9	2	1	12
MYRTLE BEACH SC	X	X	X	3	3	GALVESTON TX	X	4	1	1	6
WILMINGTON NC	X	X	X	2	2	FREEPORT TX	X	1	1	X	2
CEDAR KEY FL	X	X	1	1	2	GULF 29N 85W	1	3	3	1	8
ST MARKS FL	X	1	3	3	7	GULF 29N 87W	10	8	1	X	19
APALACHICOLA FL	X	3	4	2	9	GULF 28N 89W	39	X	X	X	39
PANAMA CITY FL	X	5	5	1	11	GULF 28N 91W	26	2	X	X	28
PENSACOLA FL	1	12	5	1	19	GULF 28N 93W	6	5	X	1	12

COLUMN DEFINITION PROBABILITIES IN PERCENT
A IS PROBABILITY FROM NOW TO 1AM MON
FOLLOWING ARE ADDITIONAL PROBABILITIES
B FROM 1AM MON TO 1PM MON
C FROM 1PM MON TO 1AM TUE
D FROM 1AM TUE TO 1AM WED
E IS TOTAL PROBABILITY FROM NOW TO 1AM WED
X MEANS LESS THAN ONE PERCENT

FORECASTER KNABB

Figure 5. Strike-Probability forecast for Katrina (from NHC, http://www.nhc.noaa.gov/archive/2005/prb/al122005.prblty.021.shtml).

2. New NHC Methods to Convey Uncertainty in the Forecast

In recent years, probabilistic forecasting has gained increasing acceptance in the meteorological community as a supplement to deterministic forecasting. The correct use of probabilistic models along with deterministic models provides a forecaster with an opportunity to convey uncertainty in the forecast.

In the early 2000s, it was apparent that the products NHC issued to the public were becoming obsolete. Other than routinely updating the 10-year average errors, an update to the Watch/Warning "cone" forecast had not occurred since its development in the early 1980s and the strike probability product still used out-of-date statistical analysis. Newer and better statistical methods were available, which along with faster computers made the new methods cost-effective. Gross et al. (2004) proposed a new way for determining TC wind speed probabilities. This new probabilistic product would not only convey uncertainty in the track forecast, but also convey uncertainty in storm intensity and wind speed radii forecasts.

A "Monte Carlo" (MC) sampling technique was employed to meet these new requirements. The MC method consists of statistically generating a sample of random numbers from a reference distribution and observing the properties of that sample. In this case, a large sample of plausible tracks relative to a given forecast track is generated by randomly selecting from track forecast error distributions derived from a historical database of NHC Official Forecast (OFCL) track forecasts. A similar approach is used for deriving the intensity and wind radii distributions. By summing the number of times a given wind speed threshold (34, 50, 64, or 100 kt) comes within a specific grid point and dividing by the total number of MC realizations, probabilities are determined (Gross et al. 2004).

The advantage of using the MC method for the new model is that the track, intensity, and wind radii error distributions sampled often are not normally distributed or fit some assumed statistical form. Since these real error distributions are sampled directly, a Gaussian distribution is not required.

The approach in this thesis is to adjust measures of the track forecast error distributions used by the MC model conditioned on measures of the track forecast confidence. The hypothesis is: If the track forecast confidence is high (low), then is it

more appropriate for the model to draw from historic track forecast errors that were produced when past forecast confidence was high (low). To examine this hypothesis, it must be determined that the track forecast error distributions are significantly different for different levels of forecast confidence.

3. New Tropical Cyclone Surface Wind Speed Probability Products

As mentioned previously, the new probability products will convey uncertainty in the track forecast, intensity, and wind speed radii. Gross et al. (2004) developed a graphical method and format to more adequately display this information to the public.

Since this new probability program is radically different from the previous operational probability products produced by the NHC, a committee within the NWS is providing oversight for the development of new operational products from the MC model output. These products were in experimental stage during the 2005 hurricane season, with the plan of providing them to the public starting in 2006 (Knaff and DeMaria 2005).

The new graphical products using the new MC model give the cumulative probability for each point in the Atlantic basin and surrounding continents that a certain wind speed threshold will be exceeded within the given time period. For example (Figure 6), the current experimental products display cumulative probabilities for time periods from 12 h to 120 h. The probabilities are determined for wind speed thresholds consisting of 34-kt (tropical storm strength), 50-kt, and 64-kt (hurricane strength) winds.

Some advantages of the new probability products from the NHC may be demonstrated with the case of Hurricane Charley in 2004. Although Hurricane Charley was forecast to hit Tampa Bay, Florida, the storm veered and the landfall point was at Port Charlotte, Florida, which caught some people off guard. Although the Watch/Warning advisory (Figure 7) had Port Charlotte within the "cone," the public tends to focus on the forecast track, or the "black line," while those on the periphery of the cone tend to wait for any changes in the forecast before taking immediate actions. Since Hurricane Charley was traveling at a small angle relative to the west coast of Florida, a small change in trajectory caused a large change in landfall location. The problem with Hurricane Charley is that when the change in track became apparent there wasn't enough time to adequately protect properties/boats and evacuate. The result was many more dollars in damage than would have been if the public had adequate notice.

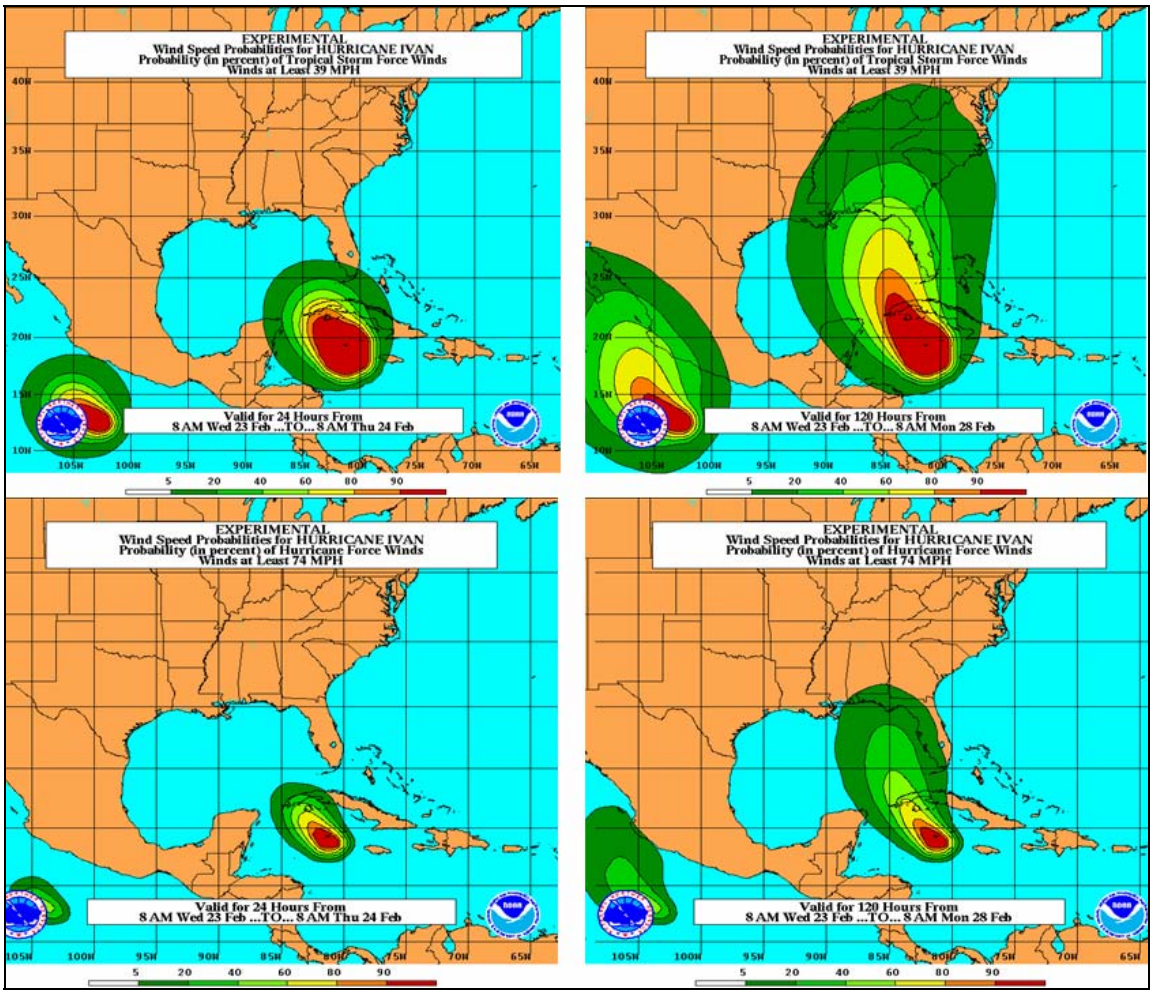

Figure 6. Experimental wind probability products for Hurricane Ivan (2004) for a) 24 h, 39 mph, b) 120 h, 39 mph, c) 24 h 74, mph, and d) 120 h, 74 mph (from NHC http://www.nhc.noaa.gov/feedback-pws-graphics2.shtml).

The new probability model (Figure 8) was in an experimental stage during Hurricane Charley. It is clear that 24 h before landfall both Tampa Bay and Port Charlotte have the same probability of hurricane force winds. If the public had the new graphic instead of the Watch/Warning "cone" graphic, perhaps more people at Port Charlotte would have been better prepared for a strike.

Examples such as this are the main motivation behind the proposed modification of the NHC probability products to include forecast confidence-conditioned error distributions addressed in this thesis. When the public receives better probabilistic information, more lives and property may be saved.

Figure 7. Five-day Hurricane Charlie Watch/Warning valid at 15 UTC 12 Aug 2004 (from Knaff and DeMaria 2005).

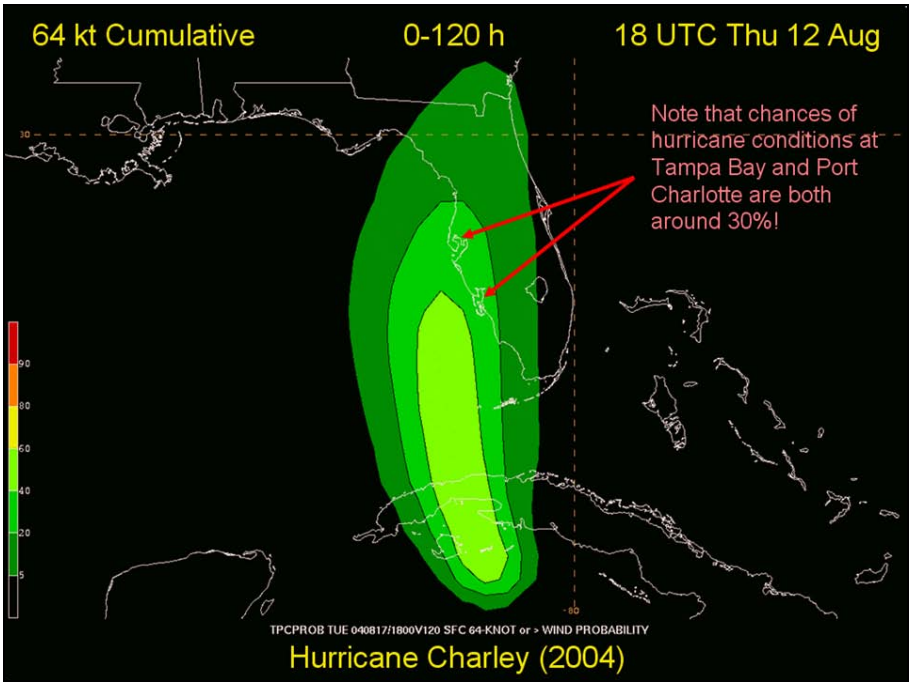

Figure 8. New graphic of cumulative probability at 64-kt winds for Charley (from NHC, http://www.nhc.noaa.gov/feedback-pws-graphics2.shtml).

B. CONSENSUS FORECASTING

1. Evolution of Consensus Forecasting in Meteorology

Consensus forecasting consists of averaging multiple predictions that were generated from slightly different starting conditions of the same model or using several different forecasts. The final objectives are to minimize bias in specific models, minimize non-predictive components, and to improve the final forecast by averaging the different forecasts. Consensus forecasting has been used in fields that have to deal with dynamic modeling, ranging from finance to biology.

Due to the lack of observations, limited computing power, different observing methods/instruments, differences in calibration among the same types of instruments, observer bias, and observer error, it is impossible to specify the true atmospheric state at any one time for input into a numerical forecast model. In addition, different models may have different biases that may consistently provide erroneous category values. The goal of consensus forecasting is to minimize the aforementioned errors by averaging many different forecasts.

Consensus forecasting in meteorology started in the late 1970s. Thompson (1977) stated the advantage in consensus forecasts:

> The purpose of this note is to draw attention to a fact that does not appear to be widely recognized or accepted, but which was probably known to Gauss in 1802. This is the incontrovertible fact that two or more inaccurate but independent predictions of the same future events may be combined in a very specific way to yield predictions that are, on the average, more accurate than either or any of them taken individually.

Thompson basically states that instead of relying on just one model to make a forecast, if we used several different models or several different initial conditions with the same model, they may be combined to produce a more accurate forecast of future conditions. Two approaches to consensus forecasting have been used to provide the future evolution of the atmosphere. A single numerical model may be integrated from many different initial conditions to provide a set of forecasts that describe the future state of the atmosphere. Alternately, the forecasts from different numerical models may be averaged. The consensus TC track forecasts use the latter method. Leslie and Fraedrich (1990) applied this method to TC track prediction by using linear combinations of forecasts from

various prediction models, and showed the consensus forecasts had a significant improvement. Goerss (2000) showed that using a consensus of operational track prediction models greatly reduced the average track forecast error over a season.

In this thesis, a product called CONU that uses this consensus approach to provide an estimate of forecast difficulty or confidence will be used to specific different forecast error distributions that should be used in the MC model.

2. CONU Product

The Consensus (CONU) product is an average of any combination of five dynamical models (GFDI, AVNI, NGPI, UKMI, and GFNI) tracks, as long as at least two tracks are available. That is, the CONU forecast is the mean of the track forecasts from the five individual models. The fact that the CONU only requires two of the five models be available to produce a track forecast makes it available to forecasters more often than other consensus track forecasts such as GUNA that requires all four model tracks to be available (see below).

As mentioned previously, consensus track forecasts almost always out-perform the individual members over time (Figure 9). The CONU product out-performs all of its individual members for every forecast interval except at 120 h, for which the UKMI did better by a few n mi. In addition, the CONU errors are almost exactly the same as for the official forecasts from the NHC, which indicates that the NHC forecasters rely on the CONU guidance and are able to add value to the CONU product at 96 h and 120 h.

An important factor for forecasters is the product availability, since it will not be used if is only available sporadically or for a limited number of cases. Goerss (2006) compared the availability of the CONU product with that of other ensembles used in TC track forecasting. The GUNS ensemble is a simple consensus of the GFDL, UKMET, and NOGAPS tracks, and GUNA is computed when track forecasts from the GFDI, UKMI, NGPI, and AVNI are all available. The CONU product is available more often than the GUNS and GUNA products with 90% or above availability for all forecast intervals (Figure 10).

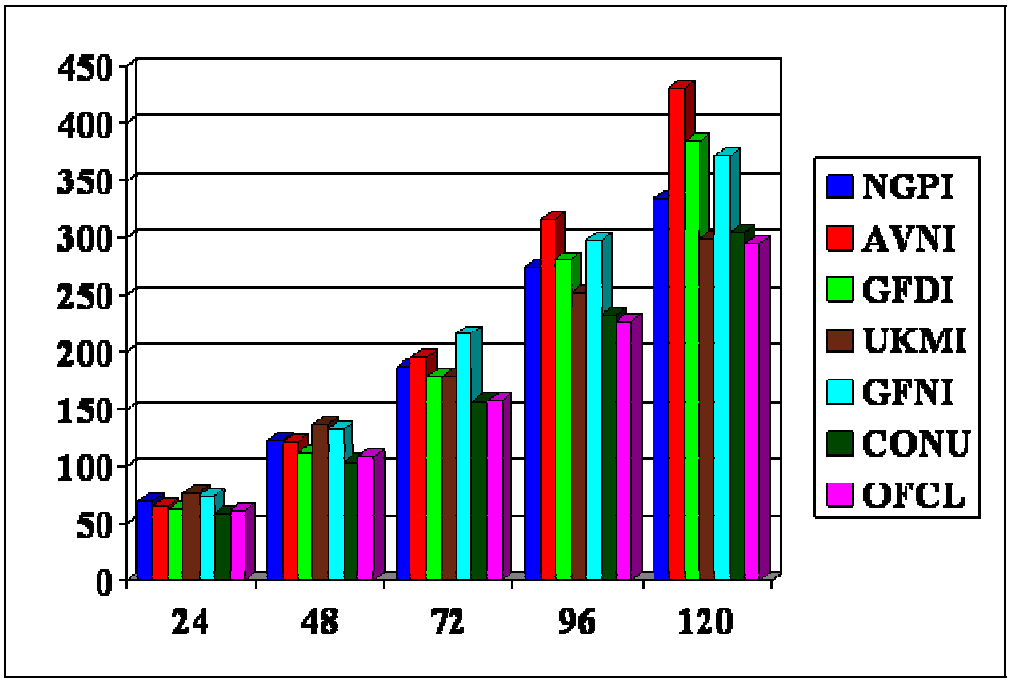

Figure 9. Non-homogeneous TC track forecast errors (n mi) on the ordinate for each forecast interval (h) displayed on the abscissa during the 2005 Atlantic season (From Goerss 2006).

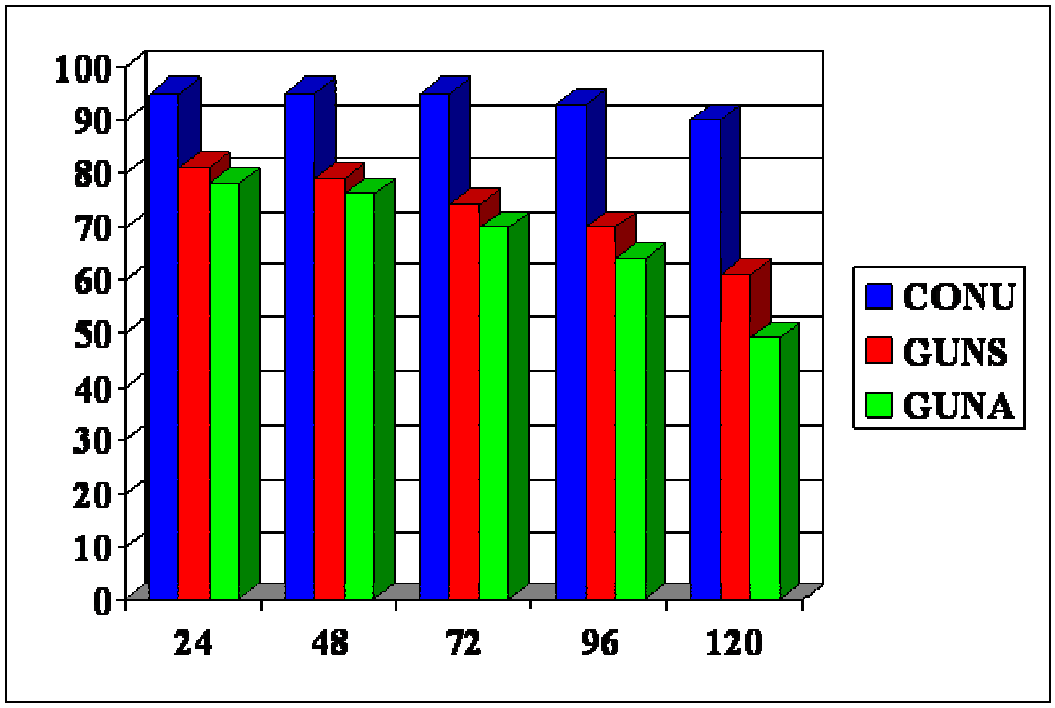

Figure 10. Availability of various consensus products during the 2005 Atlantic season. The forecast interval (h) is on the abscissa and the percent availability is along the ordinate (From Goerss 2006).

Since the CONU product is provided when only two of its five ensemble members are available and GUNA is provided only when all four dynamical model tracks are available, one would expect GUNA to have smaller track errors over time than CONU. However, Goerss (2005) found that CONU slightly outperformed GUNA out to 72 h and performed as well as GUNA out to 120 h during the 2001-2003 Atlantic hurricane seasons (Figure 11). Similar error statistics were observed for the 2005 season (Figure 12). Given the advantage of greater availability of the CONU over the GUNA (Figure 10), and the similarity in performance, CONU is the better choice for the forecaster.

As the CONU forecasting concept gained increasing acceptance in the TC community, it became clear that another potential benefit that might be gained from the CONU sample of model tracks would be to predict the error of the CONU track forecast. That is, could the spread of the model tracks be used as a measure of the confidence that the forecaster should put in today's CONU track forecast?

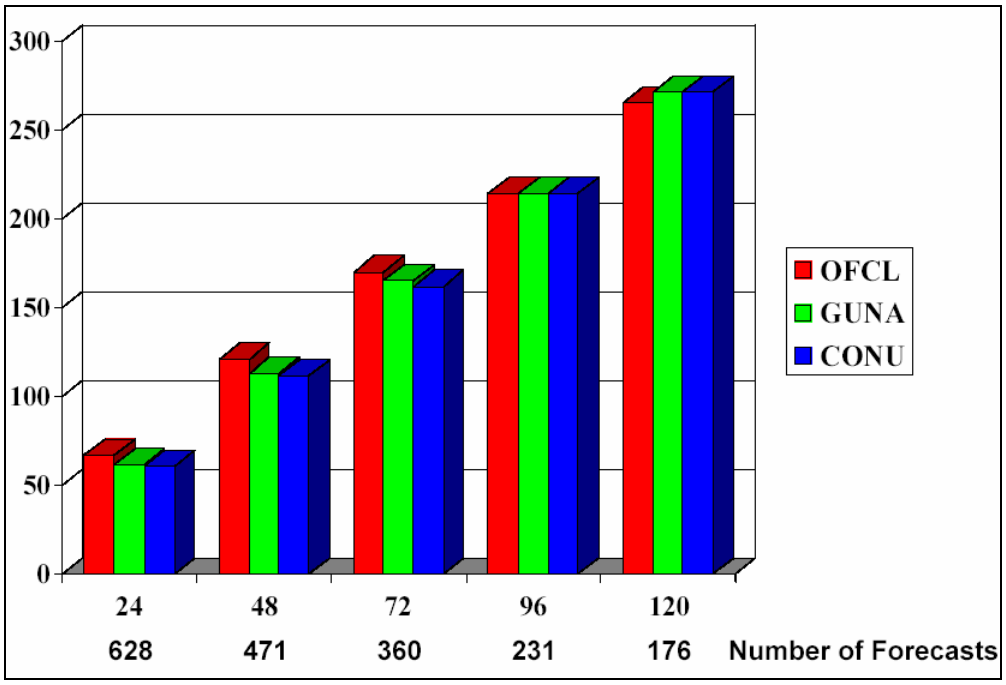

Figure 11. Homogeneous comparison of TC track forecast errors for the official, GUNA, and CONU (see insert) during the 2001-2003 Atlantic hurricane seasons. The forecast interval (h) is displayed on the abscissa and the error (n mi) is on the ordinate (From Goerss 2005).

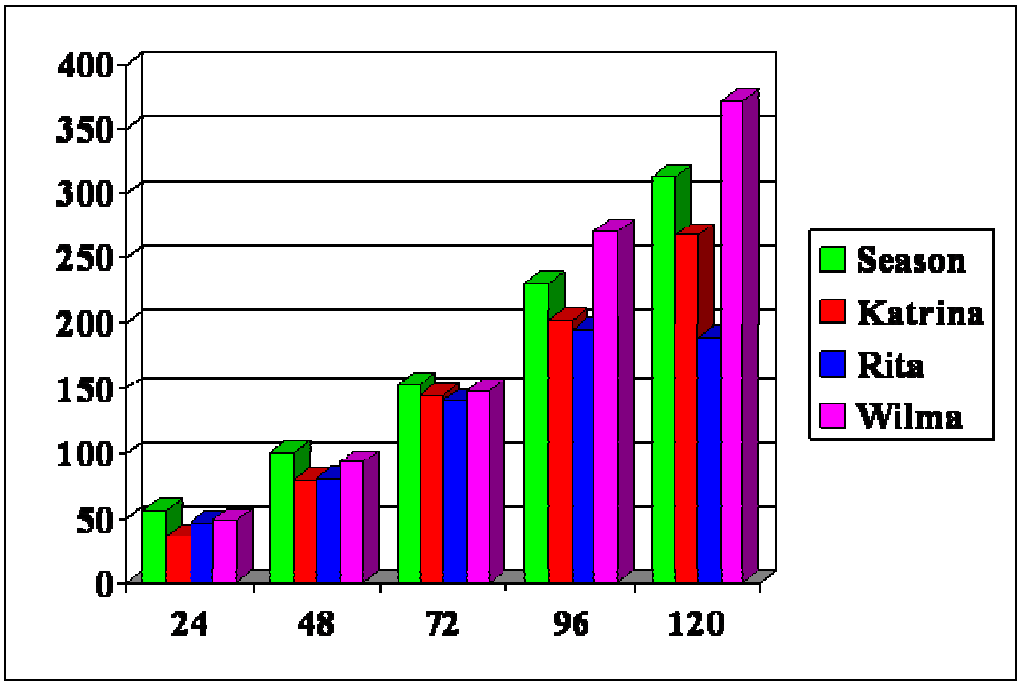

Figure 12. Track forecast errors (n mi, along ordinate) for the CONU product during the entire 2005 Atlantic season and for selected storms. The forecast interval (h) is given along the abscissa (From Goerss 2006).

3. GPCE Value

Goerss (2005) developed a method for predicting the error in the CONU TC track forecast that is called the Goerss Predicted Consensus Error (GPCE) In addition to the spread of the CONU members, other predictors are included that are available before the official forecast is issued by the NHC. These predictors include: the consensus model spread, initial and forecast TC intensity, initial TC position and forecast displacement of TC position, TC speed of motion, and the number of members available for the CONU ensemble. The predictors were compiled from the 2001 through the 2004 Atlantic hurricane seasons.

A stepwise linear regression model was derived to predict the errors in the CONU forecast, which are displayed as circular areas drawn around the CONU forecast positions. Each circle represents the area within which the TC position would verify in approximately 75% of the time. Since all of these predictors are available in real-time, the predicted error (GPCE) of the CONU forecast would be available to a forecaster before the official forecast was determined. Summed over the 2005 Atlantic hurricane season, the verifying positions were within the GPCE-derived circle approximately 75%

of the time (Figure 13). In individual storms, the verifying positions may lie outside the circle. For example, the GPCE performance for Katrina, Rita, and Wilma in Figure 13 demonstrates the variability in performance for individual storms. However, when the three storms are averaged, the GPCE circles still contain around 75%.

Some examples from Hurricane Katrina (Figure 14) demonstrate how the GPCE model works. The individual tracks represent the spread among the various members of the CONU, including the ensemble mean position which is at the center of the circles shown in Figure 14. The radius of the circle is the calculated GPCE value. The red dot represents the verifying position of Katrina.

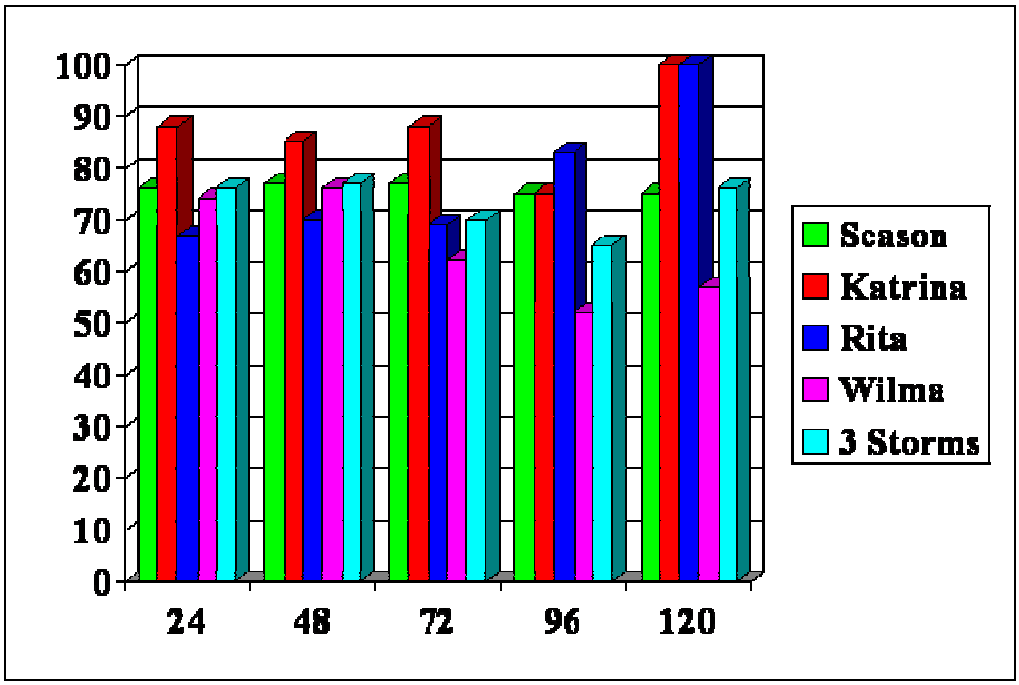

Figure 13. For each forecast interval (h, abscissa), the percent of cases (ordinate) in which the verifying position is contained within the GPCE-defined value (From Goerss 2006).

For a lead time of five days (Figure 14 a), the verifying position is just inside the GPCE circle that is centered on the CONU 120-h forecast position near Panama City, Florida. The high variability among the 120-h individual model positions that make up CONU leads to the large circle radius. For a four days lead (Figure 14 b), the ensemble members are in better agreement so the GPCE radius is smaller. Given the CONU 96-h position and the smaller circle, the verifying position falls outside the predicted forecast

error circle, which is to be expected in about one out of every four forecasts. The radius of the GPCE circle becomes progressively smaller at lead times of two days (Figure 14 d) and one day (Figure 14 e) as the ensemble member tracks are more in agreement. The verifying positions fall well within the GPCE radii for both of these forecasts.

Results from the 2005 Atlantic hurricane season (Goerss 2006) verify that the GPCE model is a reliable tool to determine forecast confidence in the CONU forecast. For the season, the verifying positions of the TCs fell within the GPCE circles 76%, 77%, 77%, 75%, and 75% for the 24-, 48-, 72-, 96-, and 120-h forecasts, respectively (Goerss 2006).

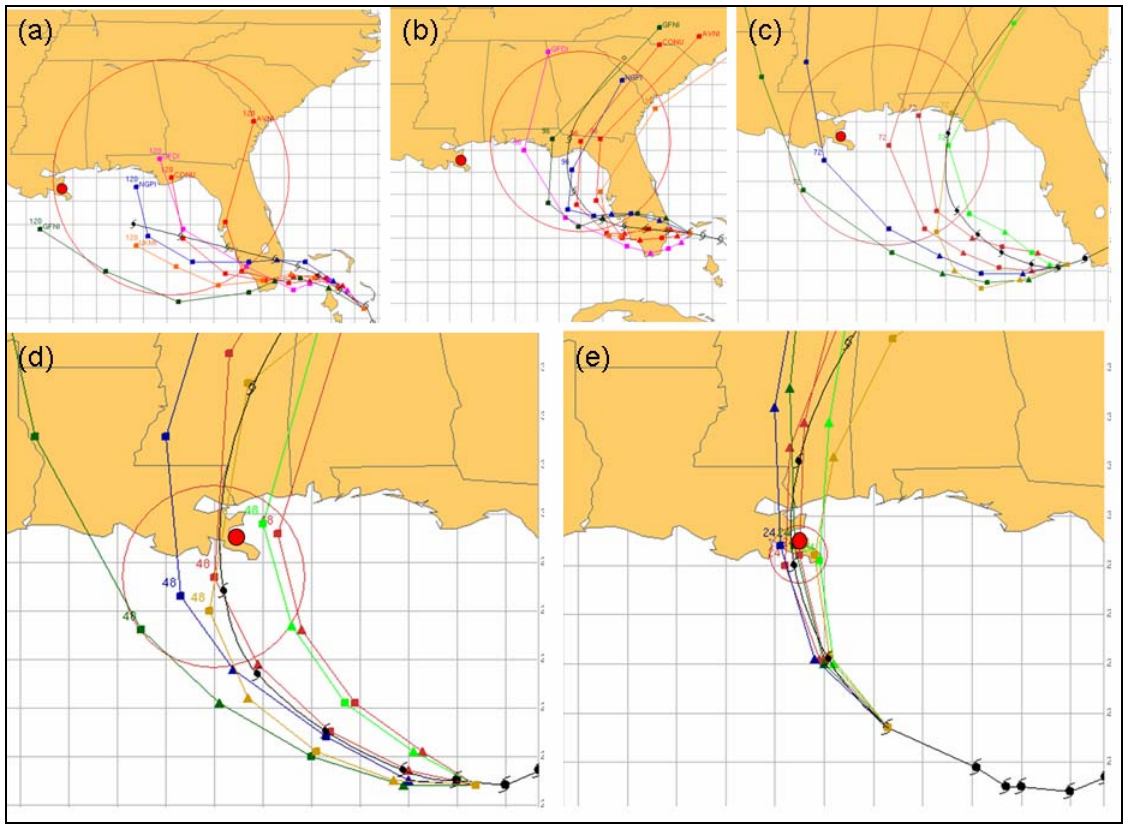

Figure 14. Predicted consensus error for CONU forecasts of Hurricane Katrina at (a) 120 h from 12 UTC 24 Aug, (b) 96 h from 12 UTC 25 Aug, (c) 72 h from 12 UTC 26 Aug, (d) 48 h from 12 UTC 27 Aug, (e) 24 h from 28 Aug 2005 (From Goerss 2006).

C. RELATIONSHIP BETWEEN MODEL SPREAD AND FORECAST SKILL

1. Measuring Forecast Position Error

The error of the forecast is determined from the verifying best-track position, which. is determined in the post-analysis stage after looking at all the available data. The error in the CONU model is dependent upon two things: 1) the mean forecast error of the individual models that make up the consensus; and 2) the degree of independence of the forecast errors of the individual models (Goerss 2000).

In addition to the magnitude of the error of the forecast track, the error may also be defined in terms of the cross-track forecast error C and the along-track forecast error A

$$E = (C^2 + A^2)^{1/2} \quad (1)$$

in which E is position error (Neumann and Pelissier 1981). The along-track forecast error represents whether the forecast was fast (positive value) or slow (negative value). The cross-track forecast error represents how far left (negative value) or right (positive value) the forecast track is relative to the verifying position (see Figure 15).

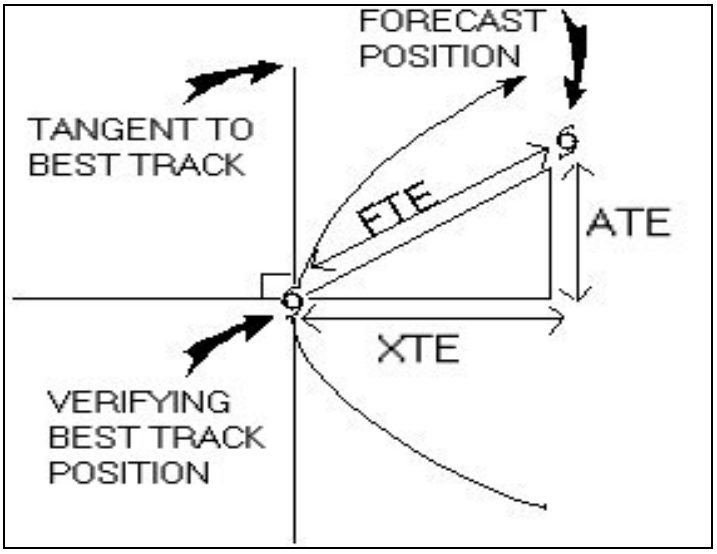

Figure 15. Definition of cross-track forecast error (XTE), along-track forecast error (ATE) and forecast track error (FTE). In this example, the forecast position is ahead of and to the right of the verifying best track position. Therefore, the XTE is positive (to the right of the best track) and the ATE is positive (ahead or faster than the best track) (from NPMOC http://www.npmoc.navy.mil/jtwc/atcr/1998atcr/ch5/chap5_page1.html.)

2. Ensemble Spread and Forecast Skill Relationship

Forecast skill is a simple measure of how different a given forecast is from the actual conditions. Given an ensemble of forecasts, the greater the spread among the members, the less confidence the forecaster may have in any specific member or in the mean of all the members. On the other hand, the better the agreement among the individual members, the higher confidence the forecaster is likely to have that the ensemble-mean forecast will adequately represent the true state of the atmosphere.

However, the correlation between ensemble spread and track forecast skill is not always so clear. For example, Goerss (2000) found no clear correlations between ensemble spread and ensemble mean error for an ensemble track prediction systems. However, Goerss found a relationship between the ensemble spread and the upper bound of error. Elsberry and Carr (2000) investigated western North Pacific TCs using a consensus of five dynamical models tracks and found that spread was a good indicator of track forecast error for low spread values. However, up to 8% of the cases with a small consensus spread had large track forecast errors. A selective consensus, in which a forecaster would eliminate only the largest track forecast error from the five models, may add value and reduce track forecast errors. Grimit and Mass (2006) stated that it is not good to simply use ensemble spread to predict a single realization of ensemble forecast track forecast error. Rather, they suggest ensemble spread should be used to correlate the distribution of ensemble-mean forecast track forecast errors over a large number of realizations.

In this thesis, the correlation between the spread in the CONU members and in the GFS ensemble members will be investigated.

3. Measuring Model Spread

As mentioned previously, the spread in the CONU model is one of the factors used to define the GPCE value. Therefore, the GPCE value is used to represent model spread for the purpose of partitioning forecast errors. When the GFS ensemble is used to measure forecast confidence, the spread among the ensemble members is defined as the average separation distance with respect to the ensemble mean.

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODOLOGY

A. DATA

1. Data Source

This thesis examines data from the 2005 Atlantic hurricane season. Although only one season, no other season has come close to the amount of activity and the amount of data collected. The 2005 Atlantic hurricane season included as many storms as two to three past seasons combined and was also well distributed across a wide geographic region. With 28 named tropical storms and an additional three unnamed tropical depressions, 648 official forecasts were issued by the NHC. Considering there are seven time periods for each forecast (12, 24, 36, 48, 72, 96, and 120 h), that means there was a potential for *4130* track forecast errors in 2005. Since a verifying position was not always available due to the TC hitting land or TC decay, the number of track forecast errors available was somewhat less. Since the 2005 Atlantic hurricane season produced such a large data set, it is expected to be large enough to test the hypothesis of this thesis.

2. Data Format

The data available from the NHC included every official forecast, a majority of the model forecasts, and all of the best-track positions of the TCs. The so-called A-Decks, B-Decks, and E-Decks from the Automated Tropical Cyclone Forecast (ATCF) system were used in this thesis. A-Decks are comprised of all of the model and ensemble forecasts available to the NHC during the season along with their OFCL forecasts. Information included in these files are the storm number, model, forecast time and period, intensity, and forecast position in longitude and latitude.

The B-Decks are the best-track positions of the Atlantic TCs in 2005. As previously mentioned, the best-track position is the verifying position of the TC after all the information has been evaluated in post-storm analysis. Included in these files are the storm number, date and time, intensity, and verifying location in longitude and latitude. Thus, the A-Decks can be used in conjunction with the B-Decks to find track forecast error.

The E-Decks contain the calculated GPCE values for the 2005 season. These files include the storm number, date and time, verifying storm location, and the calculated GPCE value. The E-Decks were used in conjunction with the A- and B-Decks to match the calculated GPCE value to the associated actual track forecast error.

B. STATISTICAL METHODS OF ANALYSIS

1. Testing for Differences in Mean

Consider the three distributions with the same means (Figure 16), but with significant differences among the three. The two distributions in the low variability case have very little overlap and thus are significantly different. Although the medium variability case has some overlap, the two distributions are distinctly different. In the high variability case, it becomes more difficult to distinguish the two populations.

In this thesis, the first goal is to demonstrate that the track forecast error distributions are significantly different when separated into the low or medium variability case demonstrated in Figure 16. If there is little difference between them as demonstrated by the high variability case in Figure 16, then most likely there will be little improvement in modifying the MC model for different variabilities since the distributions will be so similar that using them independently will not change the probability output.

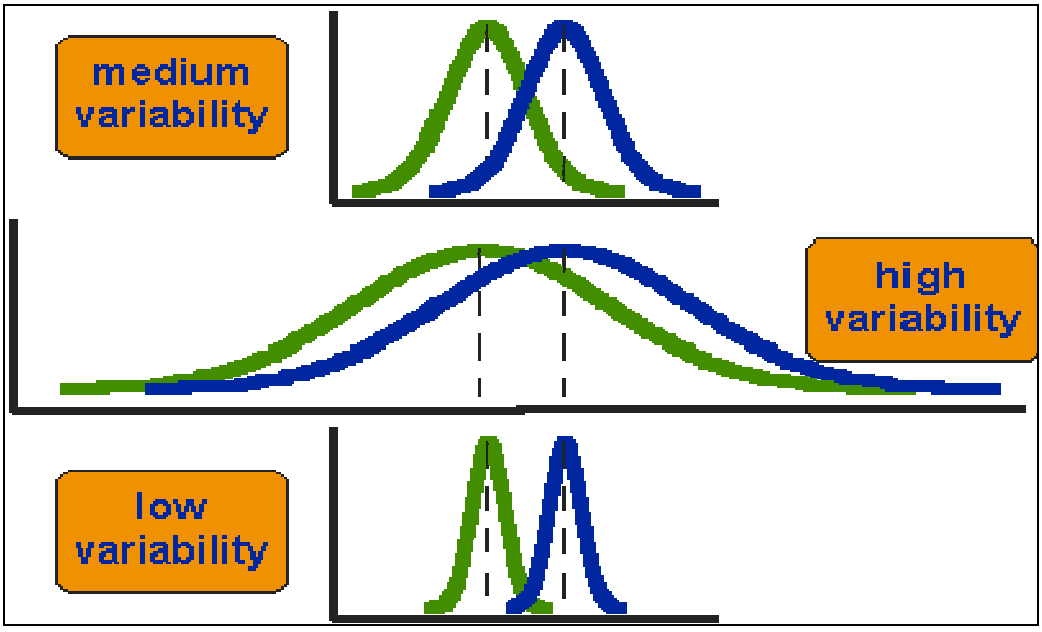

Figure 16. Three pairs of distributions with the same mean (from Trochim, http://www.socialresearchmethods.net/kb).

An objective method is used to determine if two populations are significantly different. The t-Test statistic (T) is a function of the differences between the two sample means, and takes into account the size and variance of the distributions,

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of the two samples, $\mu_0 = \mu_1 - \mu_2 = 0$ is the hypothesized difference between the two means, S_1 and S_2 are the standard deviations of the two samples, and n_1 and n_2 are the numbers of members in each sample. For this thesis, the t-statistic is evaluated using a 95% confidence level, which means the test will be in error no more then five out of 100 times.

The null hypothesis for this test is that the two means are the same ($\mu_1 - \mu_2 = 0$) (Figure 17). If the null hypothesis is true, then the t-statistic will fall in the acceptance region of the t-distribution (t-statistic < t-critical). If the null hypothesis is false, then the t-statistic will fall in the critical region of the t-distribution (t-statistic > t-critical).

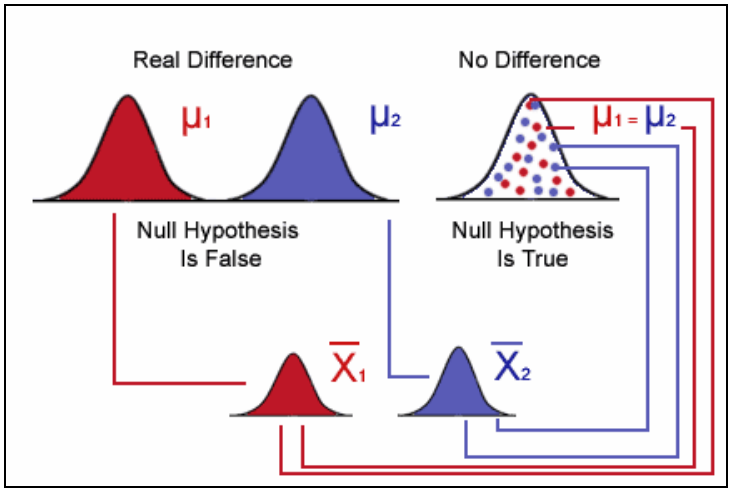

Figure 17. Hypothesis test for differences in mean (from Wadsworth, http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/ttest_betwn/ttest_betwn_02.html).

As indicated in Figure 16, the means of two samples may be the same and the variances are different. This situation may apply with cross- and along-track forecast errors since the values are both positive and negative. These errors tend to cancel at times, which leads to very little differences between the distributions. However, the means of the distributions can be used as an indicator of the skewness. Using different distributions that have the same mean but different variances also makes them separable and may also improve the MC method.

2. Testing for Differences in Variance

The test for differences in variances is simply taking the ratio of the larger sample variance over the smaller sample variance and comparing it to a F-distribution at a certain confidence level and degrees of freedom. As with the t-tests, the confidence level in this thesis will be 95%. The equation for the F-statistic is

$$F = \frac{S_2^2}{S_1^2}, \tag{3}$$

where S_2 is the larger variance.

The null hypothesis is that the two variances are equal, so if the F-statistic is smaller than the F-critical value, it lies in the acceptance region and the null hypothesis is true. If the F-statistic is larger than the F-critical value, then it is in the critical region and the null hypothesis is rejected. In other words, the variances between the two samples are significantly different.

This simple test for determining the differences in variance is an important piece of this thesis. If the Monte Carlo model were to draw from two distributions that had the same mean but different variances, then there may still be a significant improvement to the model.

3. Histograms

Some of the properties that can be determined from a histogram include the median of the data, spread of the data, skewness of the data, presence of outliers, and the presence of multiple modes. These properties can be a good indicator of a proper distribution or a random distribution. To create a histogram, the values are separated into bins of a predetermined size. The number of values in each bin determines the

frequency. The shape of the resulting histogram can give information about the skewness or variance of the distributions. In this thesis, histograms are used to compare different levels of forecast confidence for each forecast interval.

4. Linear Regressions and Correlations

Linear regression is a classical statistical method to find the relationship between the predictand (Y; dependent variable) and the predictor (X; independent variable). In this thesis, linear regression is used to examine the correlation between forecast error (Y; dependent variable) and GPCE value (X; independent variable). Three values are used in this thesis to illustrate the relationship: the multiple R which represents the correlation; the adjusted R2, which is the amount of the variance in forecast error that can be explained by the GPCE value; and P-value, which represents statistical significance of the R2 value.

A perfect correlation is not necessary when separating the distributions for the MC model. If the GPCE was perfectly correlated with the track forecast error, no probabilistic model would be needed. The expected distributions should contain some large-track forecast errors for low GPCE values and small-track forecast errors for high GPCE values to match the 75% hit rate of the GPCE radius (Figure 13).

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS AND RESULTS

A. INTRODUCTION

The goal of this thesis is to demonstrate the likely benefit of conditioning the track error distributions from which the MC model draws based on forecast confidence rather than using a single track error distribution for all forecasts. Forecast confidence is measured by the predicted error (GPCE) in the consensus track forecasts. If the forecast confidence is high, then it may be beneficial for the probabilistic model to draw from historic track errors that were produced in the past under similar forecast confidence conditions. The same logic would apply for low forecast confidence cases. The thesis goal is to determine whether these distributions are significantly different such that changing the MC model would be beneficial.

Five 2005 Atlantic hurricane season track error distributions are tested to see if they are separable in means and variance: OFCL total-track error and the corresponding GPCE value; OFCL along-track error and the corresponding GPCE value; OFCL cross-track error and the corresponding GPCE value; OFCL total-track error and the corresponding NCEP Global Forecasting System (GFS) ensemble spread; and the GFS ensemble mean total-track error and the corresponding GFS ensemble spread. For each forecast interval (12, 24, 36, 48, 72, 96, and 120 h), the track-error distributions and their corresponding measures of forecast confidence are evaluated independently.

Although the MC model draws from distributions of along- and cross-track errors, the total-track errors are examined first. If the total-track error distributions are not found to be significantly different when conditioned on forecast confidence, then there is no point to investigate further using the along- and cross-track components. Conversely, if the total-track error distributions are found to be significantly different, then an examination of the along- and cross-track errors is warranted.

The distributions of track errors were binned into terciles conditioned on forecast confidence values that are available to the forecaster when the forecast is made (i.e., GPCE value or GFS ensemble spread). The resulting track-error distributions were compared using several statistical methods: differences in means and variances (Sections IIIB.1 and 2); histogram evaluation (Section IIIB.3); and linear regression (Section

IIIB.4). However, linear regression and the differences in means were not used for the cross- and along-track errors, since the positive and negative values led to the means near zero. The key statistical test for those distributions is the differences in variances since it is the variance that impacts the probabilistic output of the MC model when it draws from the distributions of along- and cross-track errors.

Although terciles usually consist of three equal-sized distributions, the conditioned distributions for a majority of the forecast intervals in this thesis do not always consist of equal numbers of samples. Different sizes arise because the samples in the distributions (track forecast errors) were not the values used to calculate the tercile levels, but rather they correspond to the samples of GPCE values or GFS ensemble spreads. For OFCL track forecast errors conditioned on GPCE values, the distributions of the lower tercile (high forecast confidence) contain larger numbers than the other terciles for most of the forecast intervals. Many times when a GPCE value was calculated, especially for the longer forecast intervals, a corresponding OFCL track forecast error did not exist because either the TC no longer existed or became extratropical, so a verifying position could not be established. The reason the numbers of OFCL forecast track errors decrease significantly with large GPCE values at the longer forecast intervals is that many of the large GPCE values were calculated toward the end of the TC life cycle when the models usually have less skill. In other words, toward the end of a TC life cycle, the CONU ensemble members had a large spread, so the resulting GPCE value was large. Many of those large GPCE values did not have a corresponding track forecast error because the TC no longer existed at the verification time for the forecast.

For the OFCL and GFS ensemble mean total-track forecast errors conditioned on GFS ensemble spreads, terciles at each forecast interval do have roughly equal-sized distributions, which is in part due to the low number of samples from the GFS ensemble. The samples that were included in the A-Decks usually had a corresponding OFCL track forecast error.

Some of the nomenclature used in this chapter and the next may seem contradictory. Although the null-hypothesis for the test for different means and variances is that they are equal, at times the term “fail” will be used when the null test for zero

difference in means or variances is accepted, because the goal is to separate the total distribution into smaller ones with different means and/or variances. Similarly, "pass" will be used when this null test is rejected.

B. OFFICIAL TOTAL-TRACK FORECAST ERRORS CONDITIONED ON GPCE VALUE

The OFCL total-track error is the distance between the OFCL forecast position and the verifying best-track position (Figure 15). These track errors were binned in three distributions based of the corresponding tercile GPCE value. Since the GPCE value is directly correlated with the spread among members of the CONU ensemble, the GPCE values were divided into terciles to define low, average, and high forecast confidence that correspond to high, average, and low CONU ensemble spreads.

The results of the comparisons are summarized in Table 1. In this table, the characteristics of the track distributions of OFCL total forecast errors and each tercile distributions for each forecast interval are defined in terms of the sample size, mean (n mi), and standard deviation (n mi). Based on the linear regression of the OFCL total-track forecast errors on the GPCE values, the correlation (R), the amount of variance in the OFCL total-track forecast errors explained by the GPCE values (R^2), and the statistical significance (P-value) are defined in the table for each forecast interval.

Tercile track forecast error histograms (Figures 18 through 21) are compared to examine changes in skewness from high forecast confidence to low. The hypothesis is that for high forecast confidence the distributions (Figures 18a-21a) should be skewed to the right, which indicates lower forecast track errors. Next, average forecast confidence distributions (Figures 18b-21b) should be distributed about the mean to indicate a high number of mid-range track errors. Finally, low forecast confidence distributions (Figures 18c-21c) should be skewed more to the left to indicate a high number of large track errors.

Table 1. The tercile comparison table for the OFCL total-track forecast errors conditioned on GPCE values. The legend in the upper right portion of the table defines the color scheme and tercile definitions.

Forecast Interval	12-H Forecast			OFCL Total-Track Forecast Errors Conditioned on GPCE Value		
Tercile	Lower	Middle	Upper			
Samples	117	150	166	**Legend**		
Mean (n mi)	27.8	40.3	46.8	Statistically Different (at 0.05 alpha)	Within 1% C.L.of Pass/Fail	Statistically the same (at 0.05 alpha)
Standard Deviation (n mi)	18	22.7	30.9			
Total Distribution	**Samples:** 433 **Mean:** 39.4 **SD:** 26.2 **R:** 0.41 **R^2:** 0.17 **P:** 0			**Lower(L)** - Lower tercile of track forecast errors when forecast confidence was high		
Comparison	L vs. M	M vs. U	L vs. U	**Middle(M)** - Middle tercile of track forecast errors when forecast confidence was average		
Test for differences in means	**t-Stat**: 5.05 **P:** 0.00	**t-Stat**: 2.12 **P:** 0.03	**t-Stat**: 6.51 **P:** 0.00			
Test for differences in variances	**F-Stat:** 1.60 **P**: 0.00	**F-Stat:** 1.85 **P**: 0.00	**F-Stat:** 2.97 **P**: 0.00	**Upper(U)** - Upper tercile of track forecast errors when forecast confidence was low		

Forecast Interval	24-H Forecast			36-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	149	155	150	153	159	121
Mean (n mi)	46.4	60.6	78.6	63.8	88.1	107.3
Standard Deviation (n mi)	30.3	33.3	46.2	40.1	48.7	58.4
Total Distribution	**Samples:** 454 **Mean:** 61.8 **SD:** 39.4 **R:** 0.41 **R^2:** 0.17 **P:** 0			**Samples:** 433 **Mean:** 84.9 **SD:** 51.8 **R:** 0.41 **R^2:** 0.17 **P:** 0		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	**t-Stat**: 3.90 **P:** 0.00	**t-Stat**: 3.90 **P:** 0.00	**t-Stat**: 7.14 **P:** 0.00	**t-Stat**: 4.82 **P:** 0.00	**t-Stat**: 2.93 **P:** 0.00	**t-Stat**: 7.00 **P:** 0.00
Test for differences in variances	**F-Stat:** 1.20 **P**: 0.13	**F-Stat:** 1.93 **P**: 0.00	**F-Stat:** 2.32 **P**: 0.00	**F-Stat:** 1.48 **P**: 0.01	**F-Stat:** 1.43 **P**: 0.02	**F-Stat:** 2.12 **P**: 0.00

Forecast Interval	48-H Forecast			72-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	141	159	105	152	113	62
Mean (n mi)	80	115	141	124	172	191
Standard Deviation (n mi)	50.9	62.8	91.6	80.7	122	137
Total Distribution	**Samples:** 405 **Mean:** 109 **SD:** 80 **R:** 0.37 **R^2:** 0.13 **P:** 0			**Samples:** 326 **Mean:** 172 **SD:** 110 **R:** 0.29 **R^2:** 0.09 **P:** 0		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	**t-Stat**: 5.25 **P:** 0.00	**t-Stat**: 2.62 **P:** 0.01	**t-Stat**: 6.19 **P:** 0.00	**t-Stat**: 3.70 **P:** 0.00	**t-Stat**: 0.90 **P:** 0.37	**t-Stat**: 3.63 **P:** 0.00
Test for differences in variances	**F-Stat:** 1.53 **P**: 0.01	**F-Stat:** 2.12 **P**: 0.00	**F-Stat:** 3.24 **P**: 0.00	**F-Stat:** 2.27 **P**: 0.00	**F-Stat:** 1.26 **P**: 0.14	**F-Stat:** 2.97 **P**: 0.00

Forecast Interval	96-H Forecast			120-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	118	87	54	96	64	46
Mean (n mi)	156	229	293	240	254	380
Standard Deviation (n mi)	107	164	200	155	185	265
Total Distribution	**Samples:** 259 **Mean:** 209 **SD:** 159 **R:** 0.29 **R^2:** 0.08 **P:** 0			**Samples:** 206 **Mean:** 276 **SD:** 200 **R:** 0.37 **R^2:** 0.13 **P:** 0		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	**t-Stat**: 3.67 **P:** 0.00	**t-Stat**: 1.95 **P:** 0.05	**t-Stat**: 4.74 **P:** 0.00	**t-Stat**: 0.50 **P:** 0.62	**t-Stat**: 2.78 **P:** 0.01	**t-Stat**: 3.32 **P:** 0.00
Test for differences in variances	**F-Stat:** 2.34 **P**: 0.00	**F-Stat:** 1.49 **P**: 0.05	**F-Stat:** 3.47 **P**: 0.00	**F-Stat:** 1.43 **P**: 0.06	**F-Stat:** 2.04 **P**: 0.00	**F-Stat:** 2.93 **P**: 0.00

1. Analysis and Results

For the 12-h forecast interval tercile comparisons, the tests for differences in means and variances demonstrate that all three distributions are significantly different (Table 1). However, the difference in means between the middle and upper terciles were about half of the difference in means between the lower and middle, which may be a result of the short time frame of the forecast (12 h). Such a short forecast interval limits how far the forecast position can be off the verifying track. In other words, large differences in speed can result in relatively small differences in distance traveled. So even if the OFCL forecast TC speed is in error, a relatively short total-track forecast error may result. Consequently, the middle- and upper-tercile distributions look similar (Figure 18, 12-h panels b and c). Both distributions range from 10 to 120 n mi (not including outliers) with similar frequencies in each bin. Additionally, the track errors based on average forecast confidence (panel b) drop from 17 track errors to ten after 70 n mi, while the track errors based on low forecast confidence (panel c) drop from 16 to six after 80 n mi. Thus, after 70 to 80 n mi the few track errors that occurred were widely spread. The biggest difference between these two terciles is that the upper tercile has a few more outliers above 100 n mi that may account for the comparison passing both tests.

All three distributions at 12 h are skewed to the right, and the magnitude increases with decreasing forecast confidence, which is consistent with the hypothesis given above. Not many track errors are expected to skew any of the distributions toward zero values on the left. The means and variances for the 12-h forecast interval distributions increase as forecast confidence decreases (Table 1). This indicates that track errors are getting larger with more spread that indicates lowering forecast confidence.

The 24-h forecast interval tercile comparisons passed all tests except for difference in variances between the lower and middle tercile (Table 1). Although they have significantly different means and different shapes (Figure 18, 24-h panels a and b), the standard deviations differ by only 3 n mi. However, the high confidence tercile is skewed to the right, while the middle tercile distribution is somewhat centered. This difference is expected between high and average forecast confidence. Also both the error means and variances increase with decreasing forecast confidence as measured by the GPCE values.

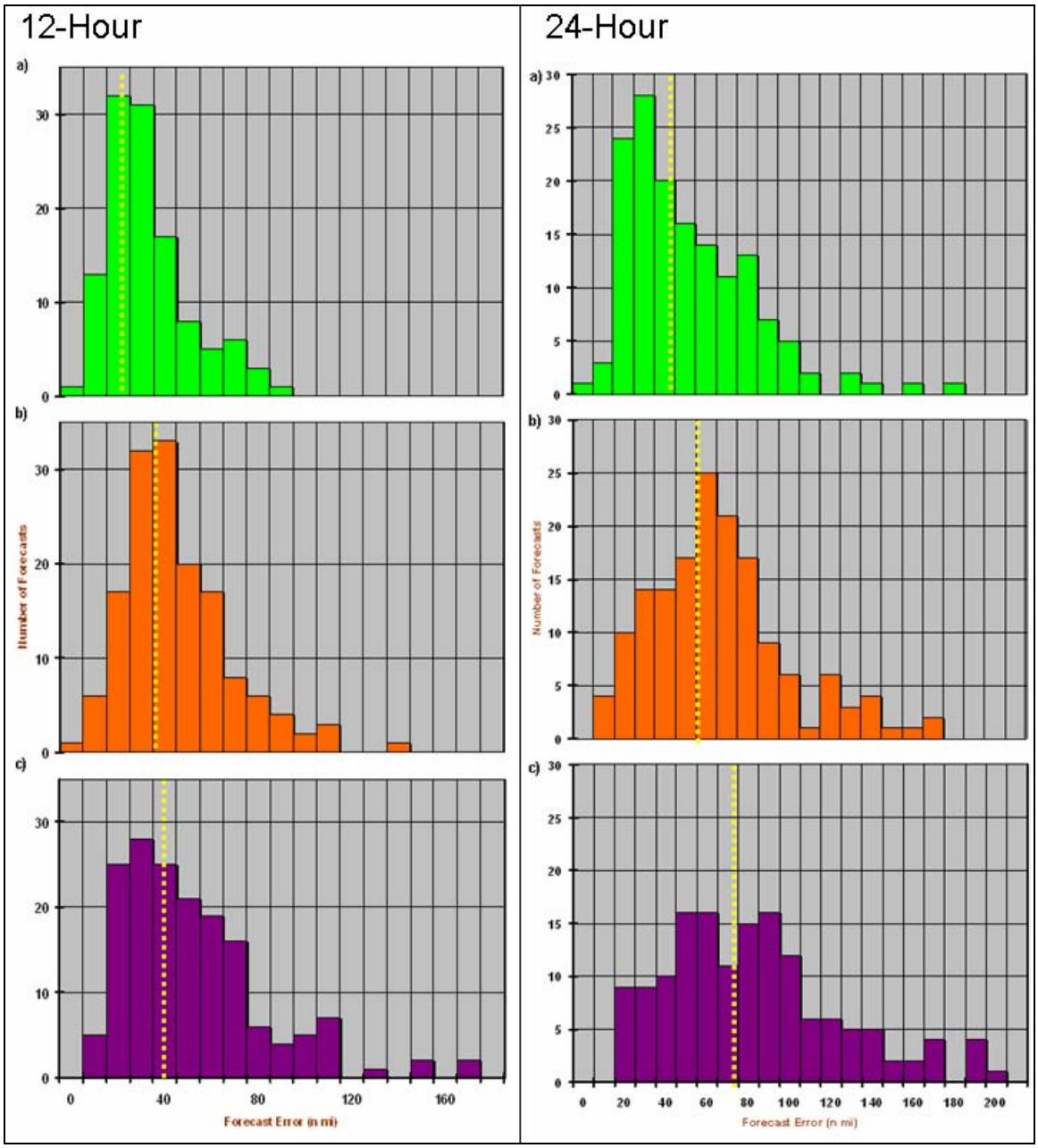

Figure 18. Histograms of 12- (left column) and 24-h (right column) OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

The 36- and 48-h forecast interval tercile comparisons both passed all statistical tests as their means and variances increase with decreasing forecast confidence (Table 1). In addition, the tercile histograms (Figure 19) show the hypothesized progression in skewness between high, average, and low forecast confidence.

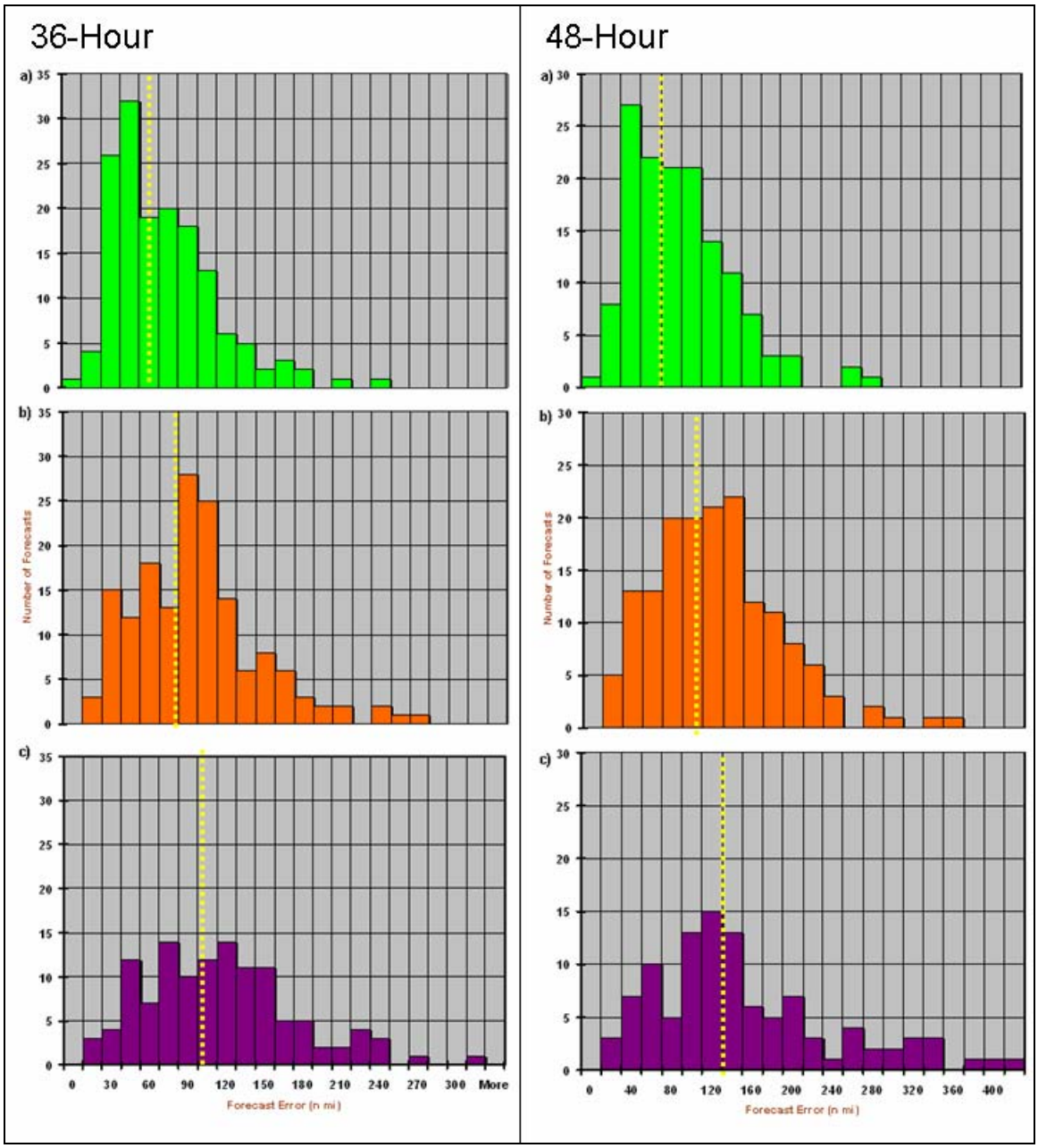

Figure 19. Histograms of 36- (left column) and 48-h (right column) OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

Both the lower- and middle-tercile and the lower- and upper-tercile comparisons passed the statistical tests for the 72- and 96-h forecast intervals (Table 1). However, the middle- and upper-tercile comparisons did not. Despite this, both the 72- and 96-h forecast intervals had increases in the means and variances with decreasing forecast confidence along with the hypothesized progression of skewness (Figure 20).

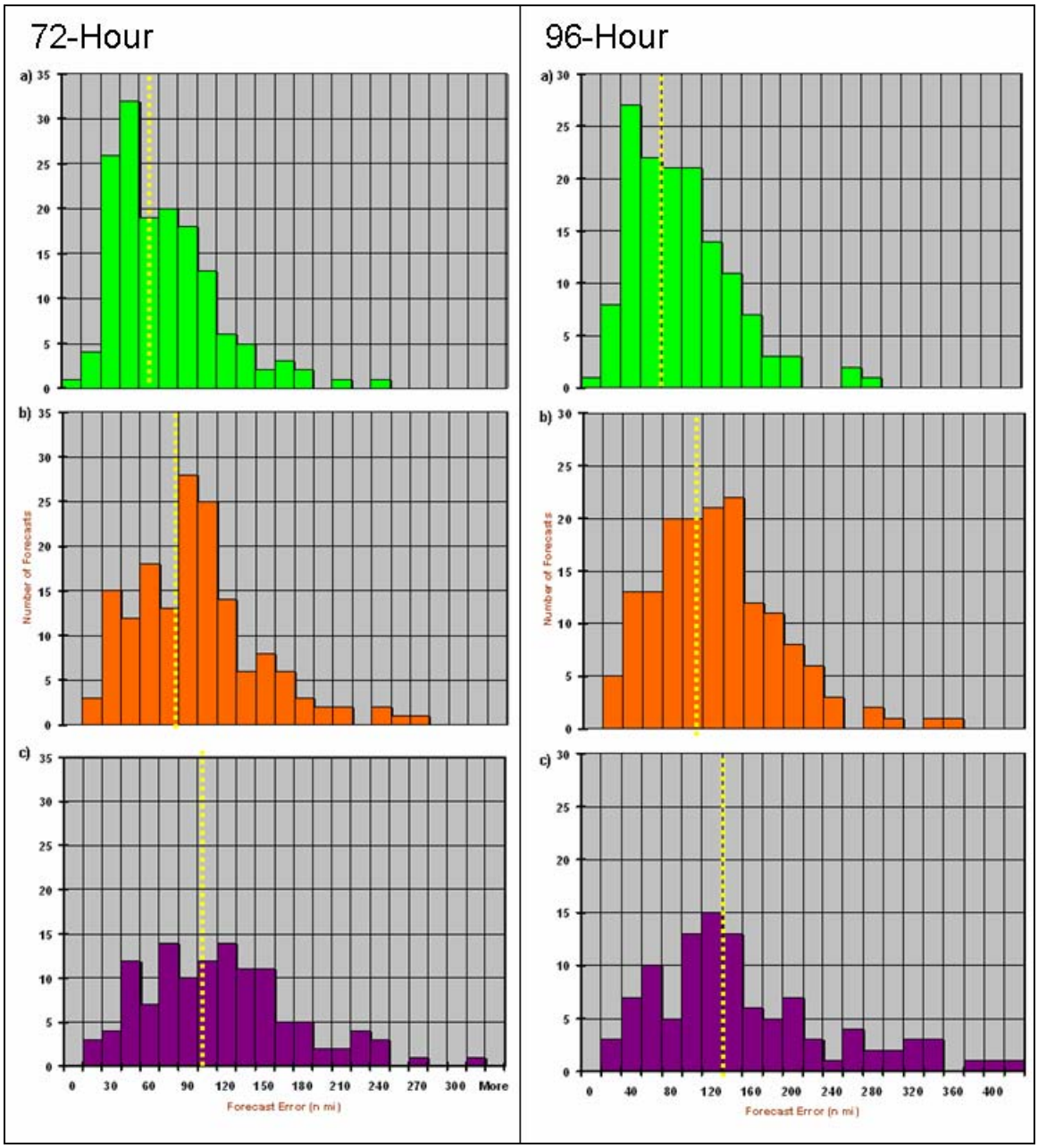

Figure 20. Histograms of 72- (left column) and 96-h (right column) OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

For the 72-h forecast middle- and upper-tercile comparison, only a 19 n mi difference occurred, while the lower- and middle-terciles had a 48 n mi difference. The corresponding difference in standard deviations was only 15 n mi, while the lower- and middle-terciles comparison had a difference of 41 n mi (Table 1). These similarities can

be seen in the widths of the distributions in Figure 20, 72-h panels b and c. Despite the failed tests, the distributions still have the hypothesized progression in skewness.

The statistical tests for the 96-h forecast interval middle- and upper-tercile comparisons both resulted in P-values of 0.05. In other words, if this were evaluated using a 96% confidence level instead of 95%, they would both fail.

Both the middle and upper and the lower- and upper-tercile comparisons passed the statistical tests for the 120-h forecast interval (Table 1). However, the lower- and middle-tercile comparisons failed both tests, although the test for differences in variances would pass at a 94% confidence level.

The reasoning for the poor results of the lower- and middle-tercile comparisons are analogous to the reasoning given for the 0.03 P-value of the middle and upper terciles test for differences in means for the 12-h forecast interval. Just as a short time frame limits the distances the TCs travel, the long time integration to 120 h means that even slow moving TCs have traveled a large distance. Consequently, small differences in speed result in large differences in distance. Therefore, the potential for larger track errors from missed forecasts is greater. Even if the OFCL forecast missed the TC speed by a small margin, a relatively large track error will result. Consequently, the lower- and middle-tercile distributions are similar (Figure 21). Both distributions have the same range (40 n mi to 680 n mi) not including outliers, and both have similar shapes.

The lower and middle tercile distributions are both skewed to the right (Figure 21). The upper tercile skewness is not easily discernable due to the small number of samples. Although the mean and variance increase as forecast confidence decreases (Table 1), the skewness of the distributions do not follow the hypothesized progression from high to low forecast confidence.

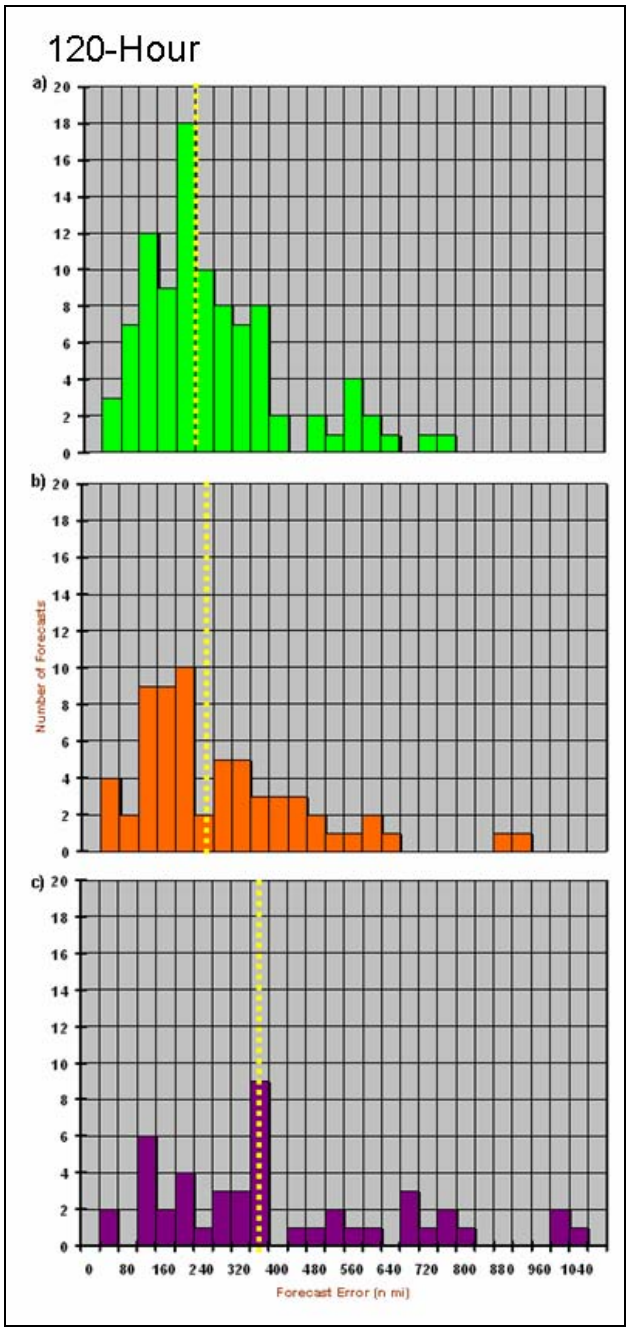

Figure 21. Histograms of 120-h OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

2. Summary

The tests for differences in OFCL forecast error means resulted in 19 of the 21 tercile comparisons having significantly different means when conditioned by the GPCE values. Similarly, the tests for differences in variances showed that 18 of the 21 tercile comparisons had significantly different variances.

The correlations between GPCE values and OFCL total-track forecast errors remained somewhat constant throughout the forecast intervals (Table 1). The 12-, 24-, 36-, 48-, and 120-h forecasts all had correlations from 0.37 to 0.41. The only exceptions were the 72- and 96-h forecasts, which both had a correlation of 0.29. Since a perfect or high correlation is not needed because uncorrelated values need to be represented in the probabilistic model, these results are consistent with expectations. These positive correlations reflect that the means and variances increased with decreasing forecast confidence for every forecast interval.

The skewness of the different terciles followed the hypothesized progression from right to left with decreasing forecast confidence for all forecast intervals except for 12 h and 120 h (Figures 18 through 21). This indicates that high forecast confidence cases are associated with more small track errors, average forecast confidence cases have more average tracks errors, and low forecast confidence cases are associated with more large track errors.

All of these factors combined indicate that the GPCE value is a good indicator of forecast confidence for the OFCL total-track forecasts. Using tercile GPCE value to condition total-track forecast error into terciles will most likely result in three significantly different distributions. Consequently, these total-track results warrant an examination of the along- and cross-track OFCL forecast errors to see if they also can be separated into significantly different distributions. If so, then those distributions may improve the probabilistic output of the MC model.

C. OFFICIAL ALONG-TRACK FORECAST ERRORS CONDITIONED ON GPCE VALUE

The OFCL along-track error is a component of the total-track error and can be used as a measure of whether the forecast is fast or slow (Figure 15). The errors can be positive (which represents a forecast that was fast) or negative (which represents a forecast that was slow). Using the same method as the previous section, the OFCL along-track forecast errors were binned into three distributions based on the corresponding tercile GPCE value.

The results of the comparison are summarized in Table 2, and the tercile comparison histograms are displayed in Figures 22 through 25. Linear regression was

not performed because the positive and negative values would be compared with only positive GPCE values. Since along-track error is a component of the total-track error, the linear regressions from the previous section are still representative for each forecast interval. Although, the tests for differences in means are displayed in Table 2, the results are not as important since the negative and positive values tend to cancel. Therefore, the values are not highlighted as pass or fail. However, the means themselves will be used to determine whether a bias exists in the OFCL along-track forecasts. The key statistical test for these distributions is the differences in variances since it is the variance that impacts the probabilistic output of the MC model when it draws from the along-track forecast error distributions.

Table 2. The tercile comparison table for the OFCL along-track forecast errors conditioned on GPCE values. The legend in the upper right portion of the table defines the color scheme and tercile definitions.

Forecast Interval	12-H Forecast			OFCL Along-Track Forecast Errors Conditioned on GPCE Value		
Tercile	Lower	Middle	Upper			
Samples	159	164	166	Legend		
Mean (n mi)	-0.5	-11.1	-18.4	Statistically Different (at 0.05 alpha)	Within 1% C.L.of Pass/Fail	Statistically the same (at 0.05 alpha)
Standard Deviation (n mi)	23.4	32.1	38.5			
Total Distribution	Samples: 489	Mean: -10.2	Standard Deviation: 32.8	Lower(L) - Lower tercile of track forecast errors when forecast confidence was high		
Comparison	L vs. M	M vs. U	L vs. U	Middle(M) - Middle tercile of track forecast errors when forecast confidence was average		
Test for differences in means	t-Stat: 3.40 P: 0.00	t-Stat: 1.86 P: 0.06	t-Stat: 5.08 P: 0.00			
Test for differences in variances	F-Stat: 1.89 P: 0.00	F-Stat: 1.44 P: 0.01	F-Stat: 2.72 P: 0.00	Upper(U) - Upper tercile of track forecast errors when forecast confidence was low		

Forecast Interval	24-H Forecast			36-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	172	152	142	167	153	112
Mean (n mi)	-2.3	-13.5	-32	-7.2	-12.2	-36.1
Standard Deviation (n mi)	41.1	50	58.2	59.6	71.6	73.4
Total Distribution	Samples: 466	Mean: -15	Standard Deviation: 51	Samples: 432	Mean: -16.5	Standard Deviation: 68.7
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 2.19 P: 0.03	t-Stat: 2.92 P: 0.00	t-Stat: 5.12 P: 0.00	t-Stat: 0.68 P: 0.50	t-Stat: 2.64 P: 0.01	t-Stat: 3.46 P: 0.00
Test for differences in variances	F-Stat: 1.47 P: 0.01	F-Stat: 1.36 P: 0.03	F-Stat: 2.00 P: 0.00	F-Stat: 1.44 P: 0.01	F-Stat: 1.06 P: 0.36	F-Stat: 1.53 P: 0.01

Forecast Interval	48-H Forecast			72-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	150	149	91	152	104	57
Mean (n mi)	-4.4	-11.8	-33.2	-8.6	-60.9	-26
Standard Deviation (n mi)	68.3	89.2	122	108	148	150
Total Distribution	Samples: 390	Mean: -14	Standard Deviation: 91.4	Samples: 313	Mean: -29.1	Standard Deviation: 132
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 0.80 P: 0.42	t-Stat: 1.46 P: 0.15	t-Stat: 2.08 P: 0.04	t-Stat: 3.08 P: 0.00	t-Stat: 1.42 P: 0.16	t-Stat: 0.80 P: 0.43
Test for differences in variances	F-Stat: 1.71 P: 0.00	F-Stat: 1.86 P: 0.00	F-Stat: 3.17 P: 0.00	F-Stat: 1.87 P: 0.00	F-Stat: 1.02 P: 0.45	F-Stat: 1.91 P: 0.00

Forecast Interval	96-H Forecast			120-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	110	80	51	88	55	45
Mean (n mi)	-18.7	-80.9	-76.4	-41.7	-75	-41
Standard Deviation (n mi)	117	227	307	225	261	384
Total Distribution	Samples: 241	Mean: -51.6	Standard Deviation: 209	Samples: 188	Mean: -51.3	Standard Deviation: 280
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 2.24 P: 0.03	t-Stat: 0.09 P: 0.93	t-Stat: 1.30 P: 0.20	t-Stat: 0.78 P: 0.44	t-Stat: 0.51 P: 0.61	t-Stat: 0.01 P: 0.99
Test for differences in variances	F-Stat: 3.75 P: 0.00	F-Stat: 1.82 P: 0.01	F-Stat: 6.85 P: 0.00	F-Stat: 1.34 P: 0.11	F-Stat: 2.17 P: 0.00	F-Stat: 2.91 P: 0.00

1. Analysis and Results

The 12- and 24-h forecast interval tercile comparisons both passed the test for differences in variances (Table 2). In addition, the variances increased as forecast confidence decreased. However, the middle and top terciles were more similar than the lower and middle, which again may be due to the short forecast intervals as discussed in the previous section. All three tercile along-track error distributions for both of these forecast intervals have a negative bias (Table 2). Interestingly, the upper limit of positive along-track errors does not increase as forecast confidence decreases for both forecast intervals (Figure 22). Rather, the negative along-track errors increase in magnitude as forecast confidence decreases.

The 36-h forecast tercile comparisons all passed the test for differences in variances except for the middle- and upper-terciles comparison (Table 2). The difference in standard deviations between the two is only 1.8 n mi compared to 12 n mi between the lower and middle terciles. However, the variance does increase with decreasing forecast confidence for this forecast interval. All three terciles for the 36-h forecast interval have a negative bias that increases with decreasing forecast confidence (Table 2). Although the middle and upper terciles statistically have the same variance, the upper tercile is shifted slightly farther to the left than the middle (Figure 23, 36-h panels b and c). This shift indicates that the largest along-track errors are negative in value, especially when forecast confidence is low.

For the 48-h forecast interval tercile comparisons, all passed the test for differences in variances (Table 2). All three terciles have negative biases that, along with variances, increase as forecast confidence decreases (Figure 23, 48-h).

The 72-h forecast interval tercile comparisons had one failure in the tests for differences in variances, which is the middle- and upper-terciles comparison (Table 2). The standard deviation is nearly identical between the two, which is consistent with the failed test for the same terciles comparison as for the total-track errors in Table 1.

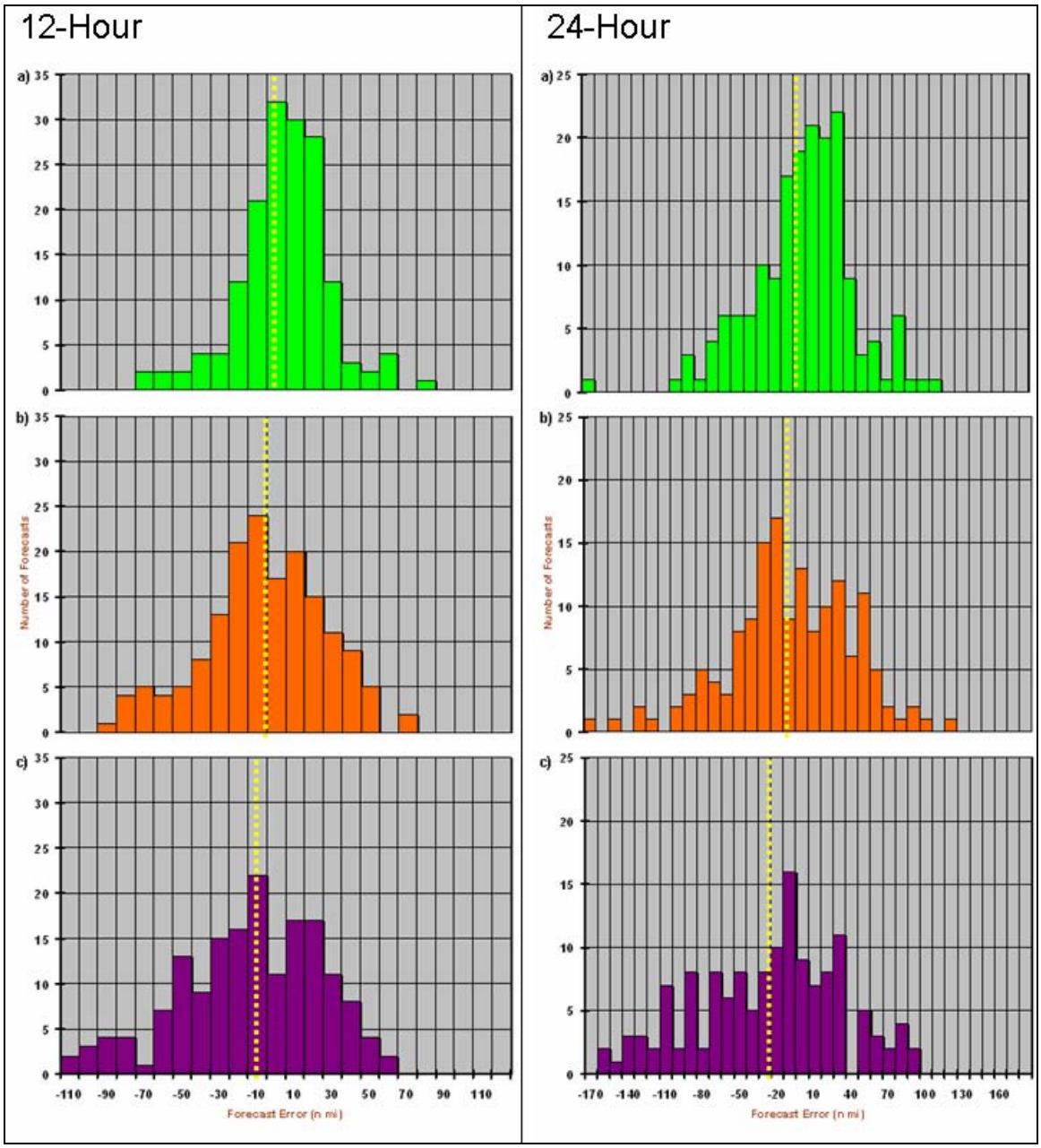

Figure 22. Histograms of 12- (left column) and 24-h (right column) OFCL along-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

The 72-, 96-, and 120-h forecast interval terciles all had negative along-track biases. For all three, the middle tercile representing average forecast confidence had the largest negative along-track bias. When taking into account the large range of along-track forecast errors at the longer forecast intervals, the differences become less

significant (Figures 24 and 25). Despite this discrepancy, variance does increase with decreasing forecast confidence for all three of these forecast intervals.

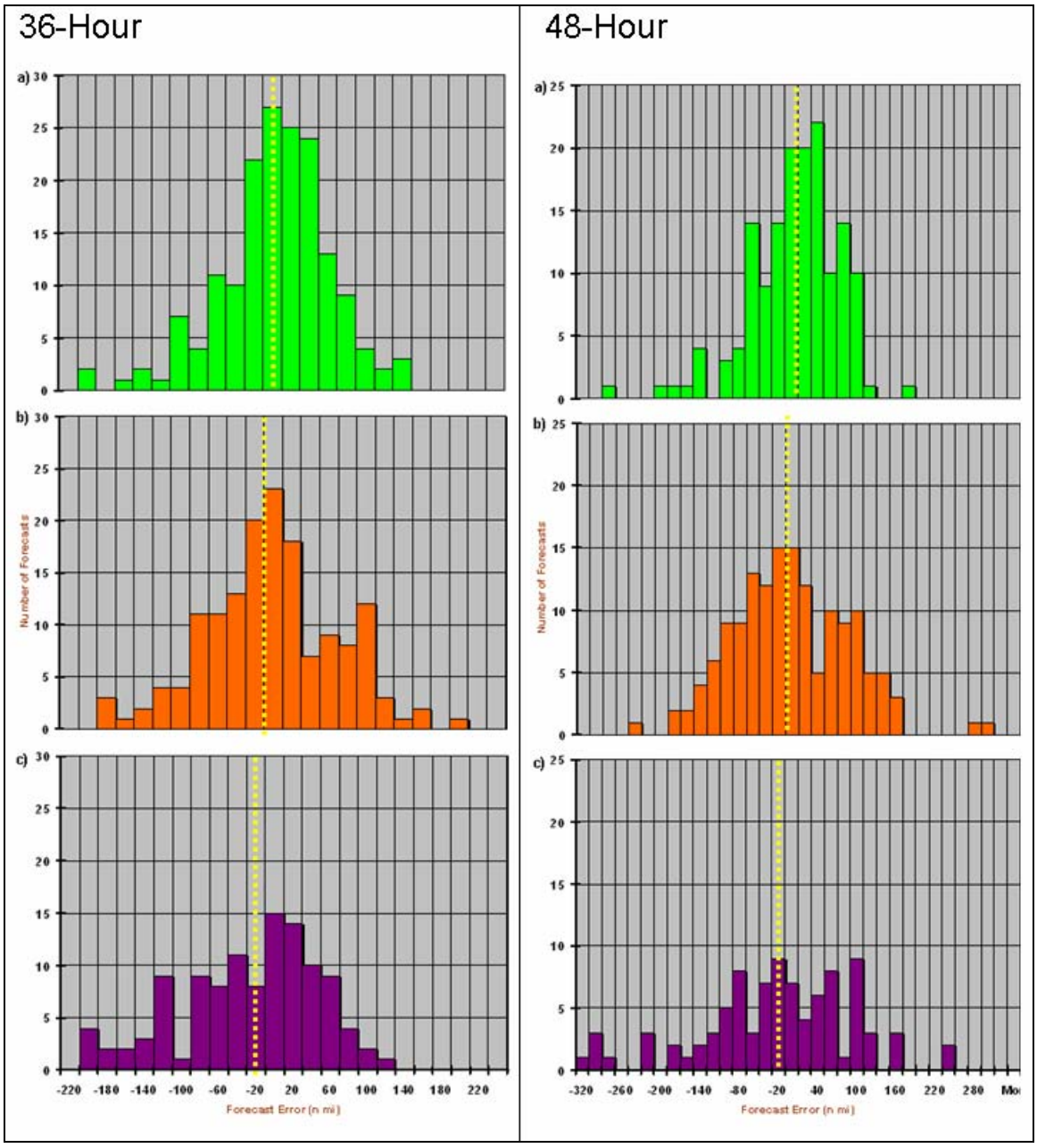

Figure 23. Histograms of 36- (left column) and 48-h (right column) OFCL along-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

The 96-h forecast tercile comparisons all passed the tests for differences in variances, while the 120-h forecast interval had one failure, the lower- and middle-terciles comparison (Table 2). This failure is for the same terciles comparison as for the total-track errors (Table 1).

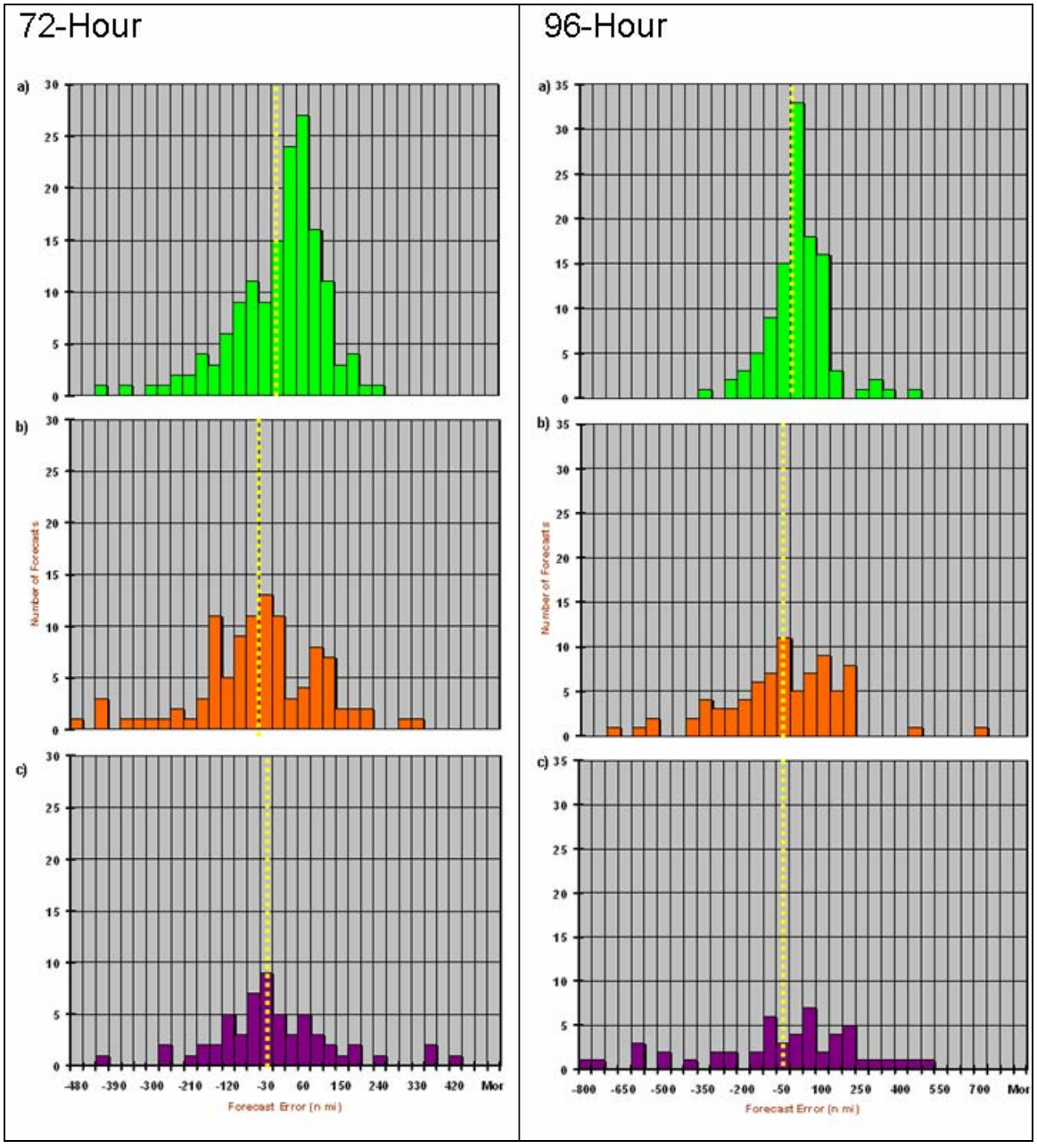

Figure 24. Histograms of 72- (left column) and 96-h (right column) OFCL along-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

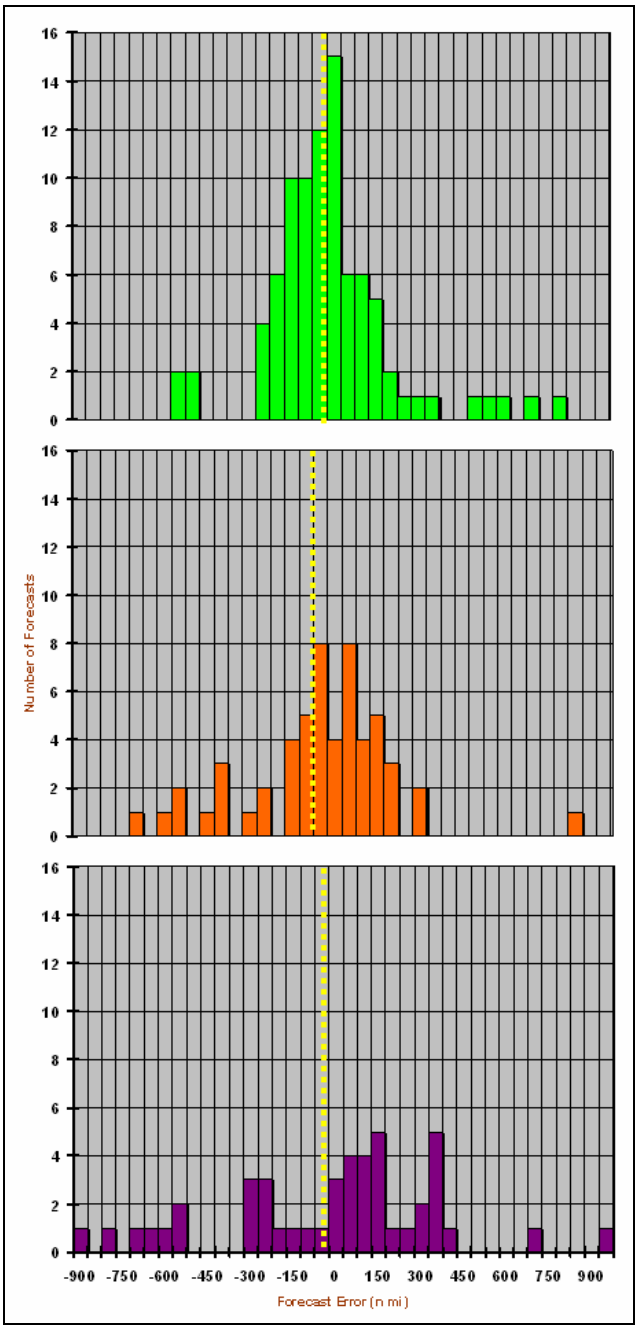

Figure 25. Histograms of 120-h OFCL along-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

2. Summary

The differences in variances tests for the OFCL along-track errors resulted in 18 of the 21 tercile comparisons having significantly different variances. Two of the failures were consistent with the total-track failures for the 72-h middle and upper terciles and 120-h lower and middle tercile comparisons. The failure of the 36-h middle- and upper-terciles along-track errors comparison did not have a corresponding failure with total-track errors. However, that comparison did have a P-value of 0.03. Conversely, the 24-h along-track error lower and middle terciles comparison passed while the corresponding total-track error terciles comparison failed.

Every along-track error tercile had a negative bias for each forecast interval, which indicates that the OFCL forecast is consistently slow. If this tendency stays consistent from year to year, the probabilistic model will account for this bias in its output.

All forecast intervals had increasing along-track error variances with decreasing forecast confidence. That is, as forecast confidence lowers, the forecast along-track error becomes less predictable. If the MC model drew from three different distributions of along-track errors based on forecast confidence, the area covered by each probability interval will increase with decreasing forecast confidence.

Taking these factors into account along with the results of the total-track error comparisons, it is clear that OFCL along-track forecast errors can be successfully stratified by forecast confidence based on the GPCE value. The new probabilistic model will most likely benefit from adopting this approach.

D. OFFICIAL CROSS-TRACK FORECAST ERRORS CONDITIONED ON GPCE VALUE

The OFCL cross-track error is a component of the total-track error that can be used as a measure of whether the forecast is to the left or right of the verifying position (Figure 15). The errors can be positive (which represents a forecast that was to the left) or negative (which represents a forecast that was to the right). Using the same method as the previous sections, the OFCL cross-track forecast errors were binned into three distributions based on the corresponding tercile GPCE value.

The results of the comparisons are presented in Table 3, and the tercile comparison histograms are displayed in Figures 26 through 29. Linear regression was not performed on the cross-track errors for the same reason it was not performed on the along-track errors. Although the tests for differences in means are displayed in Table 3, the results are not as important since the negative and positive values tend to cancel. Therefore, the values are not highlighted as pass or fail in Table 3. However, the means will be used to determine whether a bias exists in the OFCL cross-track forecasts. The key statistical test for these distributions is the differences in variances, since it is the variance that impacts the probabilistic output of the MC model when it draws from the cross-track forecast error distributions.

Cross-track errors for the OFCL forecasts are usually smaller in magnitude than the along-track errors. Therefore, the range of the distributions will be smaller than for the along-track distributions. Because of this smaller variability, some of the tests for differences in cross-track error variances have a slightly higher P-value than for the cross-track results.

Table 3. The tercile comparison table for the OFCL cross-track forecast errors conditioned on GPCE values. The legend in the upper right portion of the table defines the color scheme and tercile definitions.

Forecast Interval	12-H Forecast			OFCL Cross-Track Forecast Errors Conditioned on GPCE Value		
Tercile	Lower	Middle	Upper			
Samples	157	159	154	Legend		
Mean (n mi)	-3.1	-1.9	-5.5	Statistically Different (at 0.05 alpha)	Within 1% C.L.of Pass/Fail	Statistically the same (at 0.05 alpha)
Standard Deviation (n mi)	18.2	28.3	28			
Total Distribution	Samples: 470	Mean: -3.5	Standard Deviation: 26.2	Lower(L) - Lower tercile of track forecast errors when forecast confidence was high		
Comparison	L vs. M	M vs. U	L vs. U	Middle(M) - Middle tercile of track forecast errors when forecast confidence was average		
Test for differences in means	t-Stat: 0.50 P: 0.65	t-Stat: 1.14 P: 0.26	t-Stat: 0.90 P: 0.37			
Test for differences in variances	F-Stat: 2.41 P: 0.00	F-Stat: 1.02 P: 0.45	F-Stat: 2.37 P: 0.00	Upper(U) - Upper tercile of track forecast errors when forecast confidence was low		
Forecast Interval	24-H Forecast			36-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	161	151	140	160	153	109
Mean (n mi)	-10.3	-0.1	-13.4	-14.8	-5.6	-18.4
Standard Deviation (n mi)	32.3	48.2	52.5	45.1	68.8	79.6
Total Distribution	Samples: 452	Mean: -12.4	Standard Deviation: 50	Samples: 422	Mean: -15	Standard Deviation: 64.3
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 2.17 P: 0.03	t-Stat: 2.24 P: 0.03	t-Stat: 0.60 P: 0.55	t-Stat: 1.38 P: 0.17	t-Stat: 1.35 P: 0.18	t-Stat: 0.43 P: 0.67
Test for differences in variances	F-Stat: 2.23 P: 0.00	F-Stat: 1.19 P: 0.15	F-Stat: 2.65 P: 0.00	F-Stat: 2.32 P: 0.00	F-Stat: 1.34 P: 0.05	F-Stat: 3.11 P: 0.00
Forecast Interval	48-H Forecast			72-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	154	147	89	152	103	57
Mean (n mi)	-11.7	-12.6	-21.5	-6	-53.7	-5.6
Standard Deviation (n mi)	69.8	94.3	96.2	97	137	169
Total Distribution	Samples: 390	Mean: -14.2	Standard Deviation: 85.8	Samples: 312	Mean: -19.6	Standard Deviation: 128
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 0.09 P: 0.93	t-Stat: 0.69 P: 0.49	t-Stat: 0.84 P: 0.40	t-Stat: 3.06 P: 0.00	t-Stat: 2.27 P: 0.03	t-Stat: 0.49 P: 0.63
Test for differences in variances	F-Stat: 1.82 P: 0.00	F-Stat: 1.04 P: .041	F-Stat: 1.90 P: 0.00	F-Stat: 1.99 P: 0.00	F-Stat: 1.53 P: 0.03	F-Stat: 3.04 P: 0.00
Forecast Interval	96-H Forecast			120-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	110	80	50	89	54	44
Mean (n mi)	14.4	-1.8	29.3	32.8	24.3	83.3
Standard Deviation (n mi)	132	171	196	177	194	269
Total Distribution	Samples: 240	Mean: 12.1	Standard Deviation: 160	Samples: 187	Mean: 43.7	Standard Deviation: 207
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 0.71 P: 0.48	t-Stat: 0.93 P: 0.36	t-Stat: 0.49 P: 0.63	t-Stat: 0.26 P: 0.79	t-Stat: 1.34 P: 0.18	t-Stat: 1.27 P: 0.21
Test for differences in variances	F-Stat: 1.68 P: 0.01	F-Stat: 1.32 P: 0.14	F-Stat: 2.21 P: 0.00	F-Stat: 1.19 P: 0.23	F-Stat: 1.93 P: 0.01	F-Stat: 2.30 P: 0.00

1. Analysis and Results

The 12-, 24-, 36-, and 48-h forecast interval tercile comparisons all had similar results from the tests for differences in variances. That is, the lower- and middle- tercile and the lower- and upper-tercile comparisons passed, while the middle- and upper-tercile comparison failed except for the 36-h forecast interval that barely passed with a P-value of 0.5 (Table 3). The standard deviations of the middle and upper terciles were nearly identical for the 12- and 48-h forecasts, while the 24- and 36-h forecasts had a much smaller difference between the middle and upper terciles than they had for the lower- and middle-tercile comparisons. The similarities between the middle and upper terciles for the forecast intervals between 24 h and 48 h can be seen in Figures 26 and 27. The middle and upper terciles have nearly the same ranges, which are significantly different from the lower terciles. Along with the same ranges, the middle and upper terciles for the 24-h histograms even have roughly the same frequency distributions.

These cross-track error results are consistent with the reasoning discussed earlier regarding the short forecast intervals limiting the magnitudes of track errors. Even as forecast confidence decreases, the size of potential track error is limited, and this limit seems to be reached already when forecast confidence is average. If forecast confidence further decreases, the cross-track errors will not significantly increase. The difference with the cross-track errors from the along- and total-track errors discussed above seems to be that this line of reasoning extends out to two days instead of just one, which may be due to the relatively small magnitudes and ranges of cross-track errors when compared to along- and total-track errors. When separated by forecast confidence, the middle- and upper-tercile distributions have more similarities than the along- and total-track error middle- and upper-tercile distributions.

For all of these four forecast intervals, the cross-track error terciles have a negative bias (Table 3). For the 12-, 24-, and 36-h forecasts, the lower and upper tercile negative biases are larger than for the middle tercile, while the 48-h upper tercile bias is much larger than the biases for the middle and lower terciles. These consistent negative biases indicate the OFCL forecasts are to the left of the storm motion.

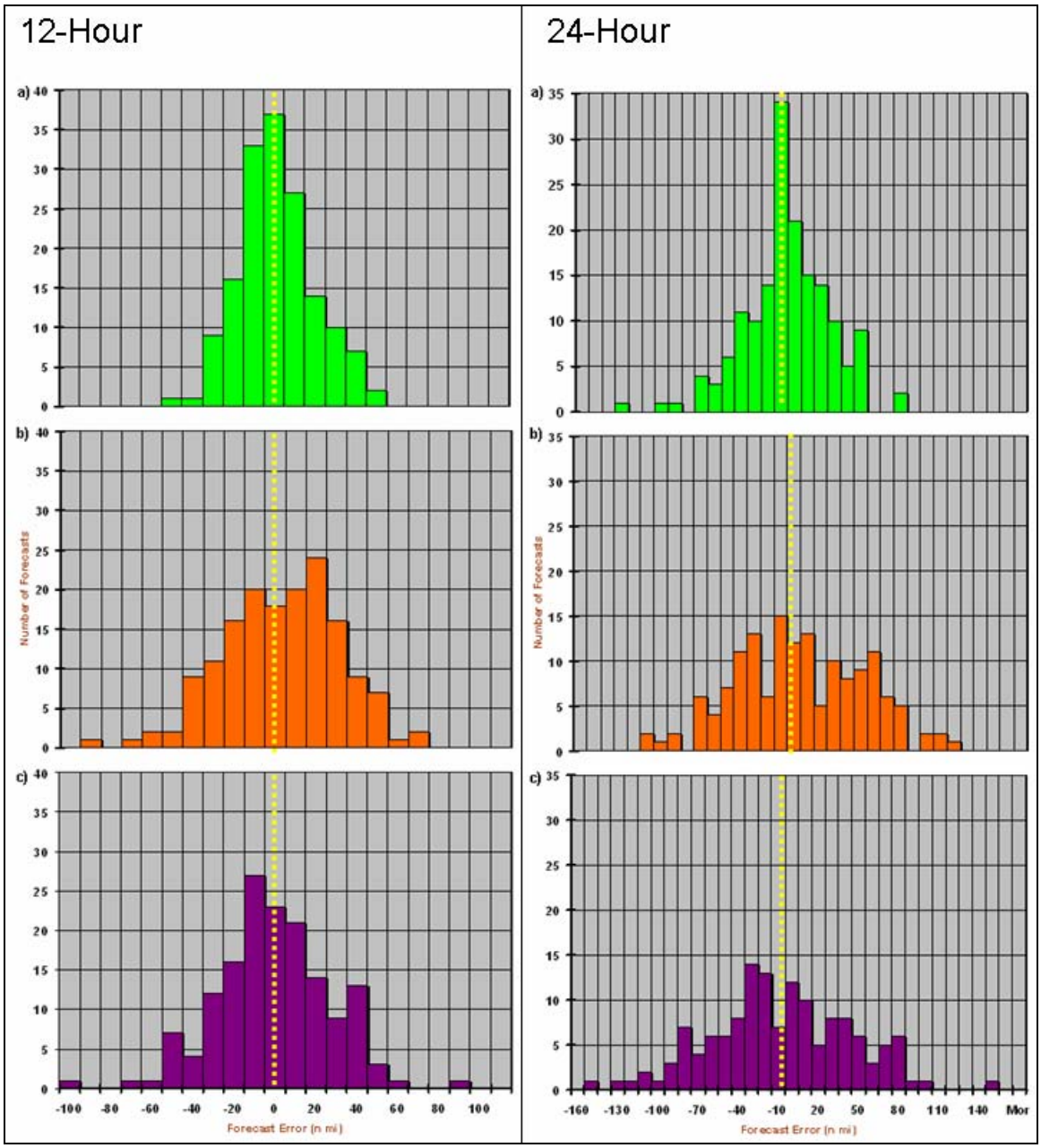

Figure 26. Histograms of 12- (left column) and 24-h (right column) OFCL cross-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

The 24-, 36- and 48-h forecast intervals all had increasing variances with decreasing forecast confidence (Table 3). However, the 12-h forecast interval had nearly identical variances for the middle and upper terciles.

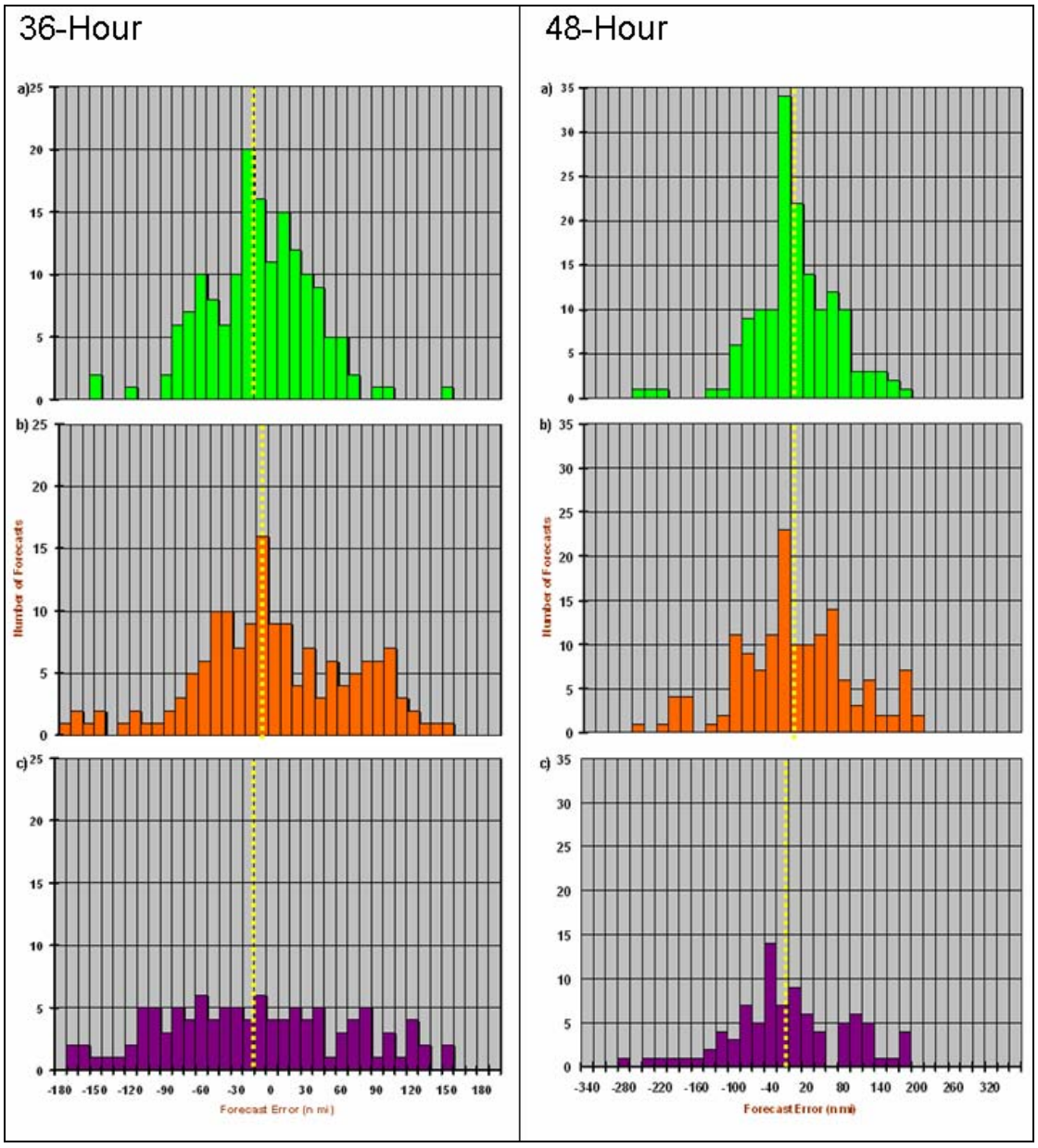

Figure 27. Histograms of 36- (left column) and 48-h (right column) OFCL cross-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

The 72-, 96-, and 120-h cross-track forecasts all had increasing variances with decreasing forecast confidence (Table 3). For all three forecast intervals, the upper tercile error distributions were shifted to the right compared to the middle tercile distribution (Figures 28 and 29). These distributions indicate that for the large forecast

intervals, the OFCL cross-track forecasts have a large bias to the right of the verifying TC positions when forecast confidence is low compared to cases with average and high forecast confidence.

The 72-h forecast interval tercile comparisons all passed the tests for differences in variances and were negatively biased (Table 3). However, the bias in the cross-track forecasts in the middle tercile was much larger than the biases for the lower and upper terciles.

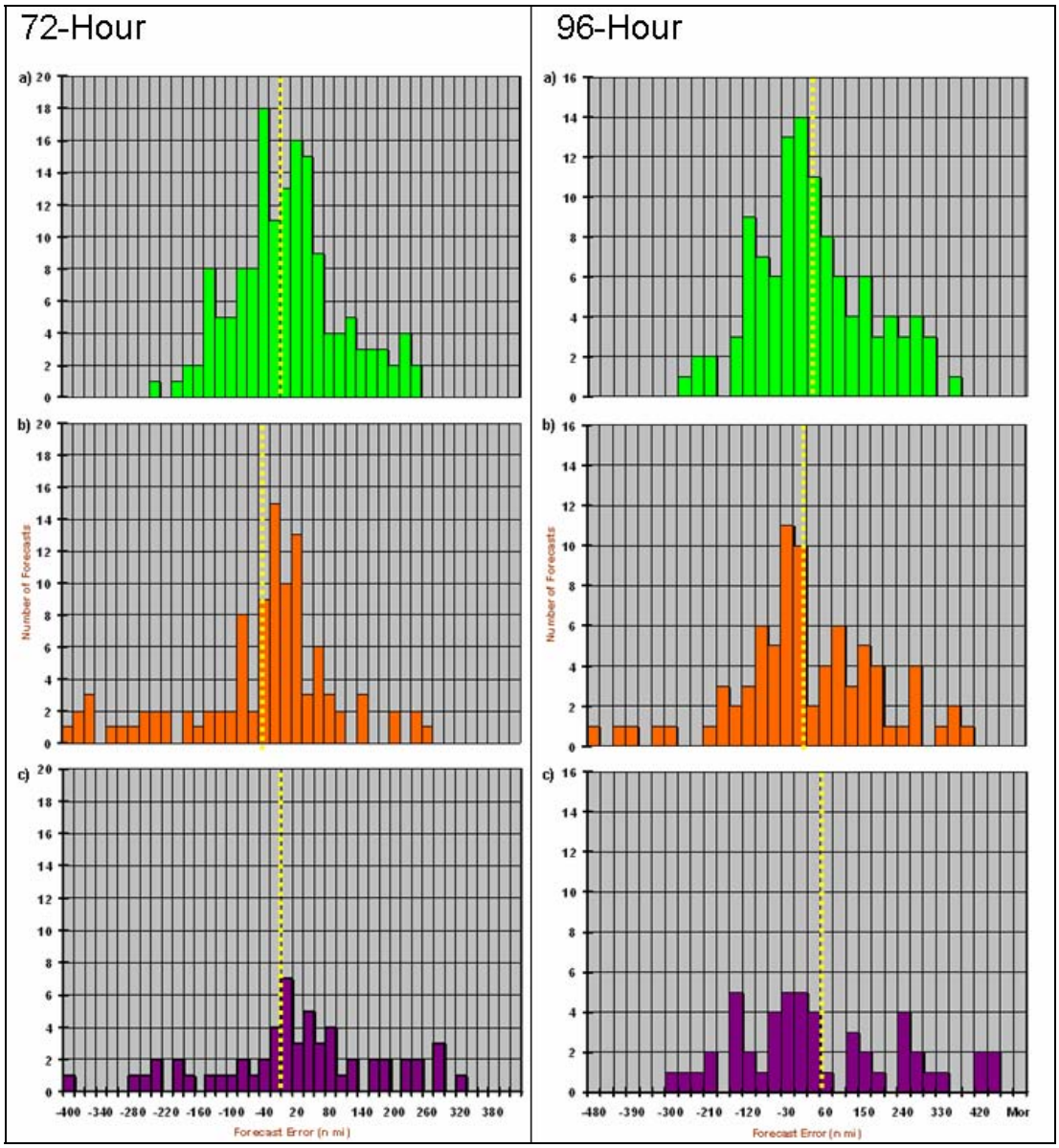

Figure 28. Histograms of 72- (left column) and 96-h (right column) OFCL cross-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

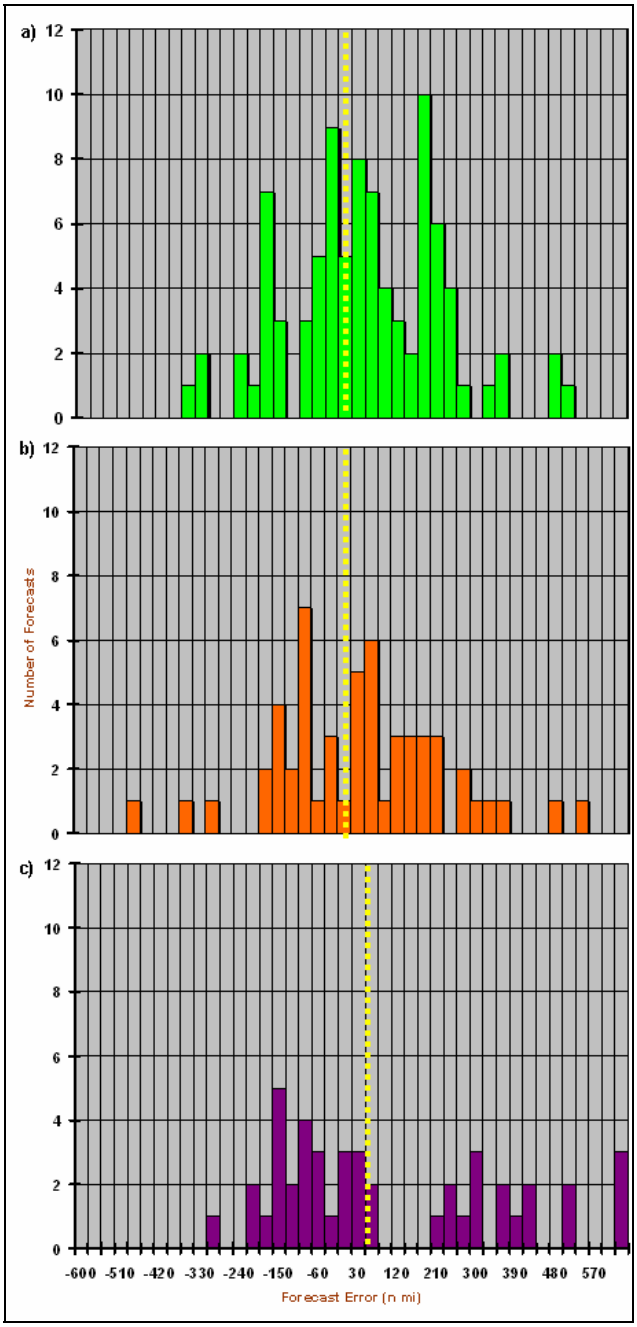

Figure 29. Histograms of 120-h OFCL cross-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

The 96-h cross-track forecast middle- and upper-tercile comparison failed the test for differences in variances, while the other two comparisons passed (Table 3). This accounts for the 0.05 P-value for the middle- and upper-tercile comparison for the 96-h total-track errors (Table 1) even though the 96-h along-track errors passed (Table 2).

For the 96- and 120-h forecasts, the cross-track forecast bias shifts from negative (left) to positive (right), which may be due to the evolution of TCs in the Atlantic basin. With few exceptions, a TC will eventually recurve to the north. At these long (96 and 120-h) forecast intervals that are likely involving the TC final stages, the movement to the north starts to increase in speed. Since the biases for these two forecast intervals are to the right, it may indicate that the OFCL forecasts tend to exaggerate this northward motion.

Consistent with the along- and total-track tercile comparisons, the 120-h middle- and upper-tercile comparison failed the test for differences in variances while the other two comparisons passed. As previously discussed, the explanation may be due to the large forecast interval. Small track errors are infrequent since even small errors in forecast motion of a TC can result in large track errors over 120 h.

2. Summary

The differences in variances test resulted in 16 of the 21 comparisons having significantly different variances (Table 3). The only cross-track error test failure in the 120-h lower- and middle-tercile comparison is consistent with both the total- and along-track errors (Table 1 and 2). The 36-h middle- and upper-tercile comparison P-value of 0.5 is consistent with the along-track error test failure for the same comparison. Interestingly, the total-track error passed the test for differences in variances for this same 36-h comparison. The 96-h cross-track error test failure for the middle- and upper-tercile comparison is consistent with the same total-track error comparison, which resulted in a P-value of 0.05. Finally, the 12- and 48-h tercile comparison failure was unique to cross-track errors.

With the exception of the 12-h forecast interval, all other forecast intervals had increasing cross-track variance with decreasing forecast confidence, which indicates that for nearly all of the forecast intervals the forecast cross-track error becomes less predictable as forecast confidence lowers. Similar to the along-track errors, if the MC model drew from three cross-track error distributions based on forecast confidence, the area covered by each probability interval will increase with decreasing forecast confidence.

The only discrepancy between the cross-track and along-track comparisons was the failures of the middle- and upper-tercile tests for differences in variances. This test failed four of the seven forecast intervals and resulted in a 0.05 P-value in another, while the lower- and middle-tercile comparisons had only one failure. These failures indicate that despite the cross-track variance increasing with decreasing forecast confidence for all except the 12-h forecast interval, the new probabilistic model will benefit much more from the separation of low from average and high forecast confidence for cross-track values than it would from the separation of the high from the low or average forecast confidence.

Considering these tests along with the results of the along- and total-track error comparisons, it is clear that OFCL track forecast errors can be successfully stratified by forecast confidence based on the GPCE values. The new probabilistic model will most likely benefit from adopting this approach.

E. OFFICIAL TOTAL-TRACK FORECAST ERRORS CONDITIONED ON GFS ENSEMBLE SPREADS

The OFCL total-track errors (equation 1) were also binned into three distributions based on the corresponding GFS ensemble spreads. The GFS ensemble spread is defined as the average distance of the ten individual members (GFS Positive One through Five and Negative One through Five) track positions from the mean track position. The GFS ensemble spread values were then divided into terciles of low, average, and large GFS ensemble spread to represent high, average, and low forecast confidence. As mentioned earlier, fewer forecast errors are in these samples due to the limited availability of the GFS ensemble in the A-Decks compared to the availability of the CONU model.

The results of these comparisons are summarized in Table 4. In addition, tercile comparison histograms (Figures 30 through 33) are compared to see if they follow the same hypothesized progression in skewness from high forecast confidence to low confidence as discussed in Section B.

Table 4. The tercile comparison table for the OFCL total-track forecast errors conditioned on GFS ensemble spreads. The legend in the upper right portion of the table defines the color scheme and tercile definitions.

Forecast Interval	12-H Forecast			OFCL Total-Track Forecast Errors Conditioned on GFS Ensemble Spread		
Tercile	Lower	Middle	Upper	Legend		
Samples	94	101	105	Statistically Different (at 0.05 alpha)	Within 1% C.L.of Pass/Fail	Statistically the same (at 0.05 alpha)
Mean (n mi)	32.1	38.1	43.2			
Standard Deviation (n mi)	25.1	25.7	22.9			
Total Distribution	Samples: 300 Mean: 38 SD: 29.4 R: 0.20 R^2: 0.04 P: 0			Lower(L) - Lower tercile of track forecast errors when forecast confidence was high		
Comparison	L vs. M	M vs. U	L vs. U	Middle(M) - Middle tercile of track forecast errors when forecast confidence was average		
Test for differences in means	t-Stat: 1.64 P: 0.10	t-Stat: 1.50 P: 0.13	t-Stat: 3.23 P: 0.00			
Test for differences in variances	F-Stat: 1.04 P: 0.42	F-Stat: 1.25 P: 0.13	F-Stat: 1.20 P: 0.18	Upper(U) - Upper tercile of track forecast errors when forecast confidence was low		

Forecast Interval	24-H Forecast			36-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	91	97	99	82	89	89
Mean (n mi)	50	64.3	68.8	74.3	86.7	93.9
Standard Deviation (n mi)	37.3	39.5	31.9	59.6	55.2	47.4
Total Distribution	Samples: 287 Mean: 61.4 SD: 37 Coorelation: 0.15, 0.02, 0.01			Samples: 260 Mean: 85.3 SD: 47.4 R: 0.12 R^2: 0.01 P: 0.04		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 2.55 P: 0.01	t-Stat: 0.88 P: 0.38	t-Stat: 3.72 P: 0.00	t-Stat: 1.44 P: 0.15	t-Stat: 0.95 P: 0.34	t-Stat: 2.37 P: 0.02
Test for differences in variances	F-Stat: 1.12 P: 0.29	F-Stat: 1.53 P: 0.02	F-Stat: 1.36 P: 0.07	F-Stat: 1.26 P: 0.15	F-Stat: 1.26 P: 0.14	F-Stat: 1.58 P: 0.02

Forecast Interval	48-H Forecast			72-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	72	72	72	51	51	50
Mean (n mi)	87.5	98.3	127	122	177	145
Standard Deviation (n mi)	52.9	67.4	77.4	70	122	113
Total Distribution	Samples: 216 Mean: 104 SD: 68.4 R: 0.20 R^2: 0.03 P: 0			Samples: 152 Mean: 148 SD: 105 R: 0.03 R^2: 0.01 P: 0.68		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 1.07 P: 0.14	t-Stat: 2.33 P: 0.02	t-Stat: 3.54 P: 0.00	t-Stat: 2.82 P: 0.01	t-Stat: 1.34 P: 0.18	t-Stat: 1.27 P: 0.21
Test for differences in variances	F-Stat: 1.63 P: 0.02	F-Stat: 1.32 P: 0.12	F-Stat: 2.14 P: 0.00	F-Stat: 3.32 P: 0.00	F-Stat: 1.16 P: 0.30	F-Stat: 2.86 P: 0.00

Forecast Interval	96-H Forecast			120-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	32	31	32	18	18	18
Mean (n mi)	184	241	250	294	199	426
Standard Deviation (n mi)	96	150	196	216	98	282
Total Distribution	Samples: 95 Mean: 306 SD: 154 R: 0.15 R^2: 0.01 P: 0.14			Samples: 54 Mean: 306 SD: 229 R: 0.36 R^2: 0.12 P: 0.01		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 1.79 P: 0.08	t-Stat: 0.21 P: 0.84	t-Stat: 1.72 P: 0.09	t-Stat: 1.69 P: 0.10	t-Stat: 3.21 P: 0.00	t-Stat: 1.57 P: 0.13
Test for differences in variances	F-Stat: 2.45 P: 0.01	F-Stat: 1.70 P: 0.07	F-Stat: 4.16 P: 0.00	F-Stat: 4.82 P: 0.00	F-Stat: 8.26 P: 0.00	F-Stat: 1.71 P: 0.14

1. Analysis and Results

The 12-h forecast tercile comparisons only passed one test, which was the lower- and upper-terciles difference in means (Table 4). The lower and middle terciles are nearly identical in means, standard deviations, shapes, and skewness (Figure 30, 12-h panels a, b, and c). The upper tercile distribution differs in mean total-track errors from the lower tercile and is more skewed than both the lower and middle terciles. However, the skewness does not follow the hypothesized progression from high to low forecast confidence as all three terciles are skewed to the right.

For the 24-h forecast tercile comparisons, the middle- and upper-tercile comparison failed the test for difference in means, while the lower- and middle-tercile and the lower- and upper-tercile comparisons failed the difference in variances test (Table 4). In this forecast interval, the skewness does follow the hypothesized progression from high to low forecast confidence (Figure 30, 24-h panels a to c).

The tests for the 36-h forecast interval resulted in the lower- and upper-tercile comparison tests passing for difference in means and difference in variances (Table 4). However, the other comparisons failed both tests. Whereas the high confidence tercile is skewed to the right, the middle and low confidence terciles are also skewed to the right but to a lesser extent (Figure 31, 36-h panels a to c). The middle and upper terciles do not follow the hypothesized progression from high to low forecast confidence.

The 12-, 24-, and 36-h forecast interval means increase with decreasing forecast confidence (Figures 31 and 32), but the variances do not (Table 4). For each forecast interval, the upper tercile variance is less than the middle- and lower-tercile variances.

The 48-h forecast interval had the best results of the seven forecast intervals. Only the lower- and middle-terciles comparison test for difference in means and the middle- and upper-terciles test for difference in variances failed. However, all of the tercile distributions are nearly symmetric and thus do not follow the hypothetical progression (Figure 32, 48-h panels a to c). The 48-h forecast error means and variances do increase with decreasing forecast confidence.

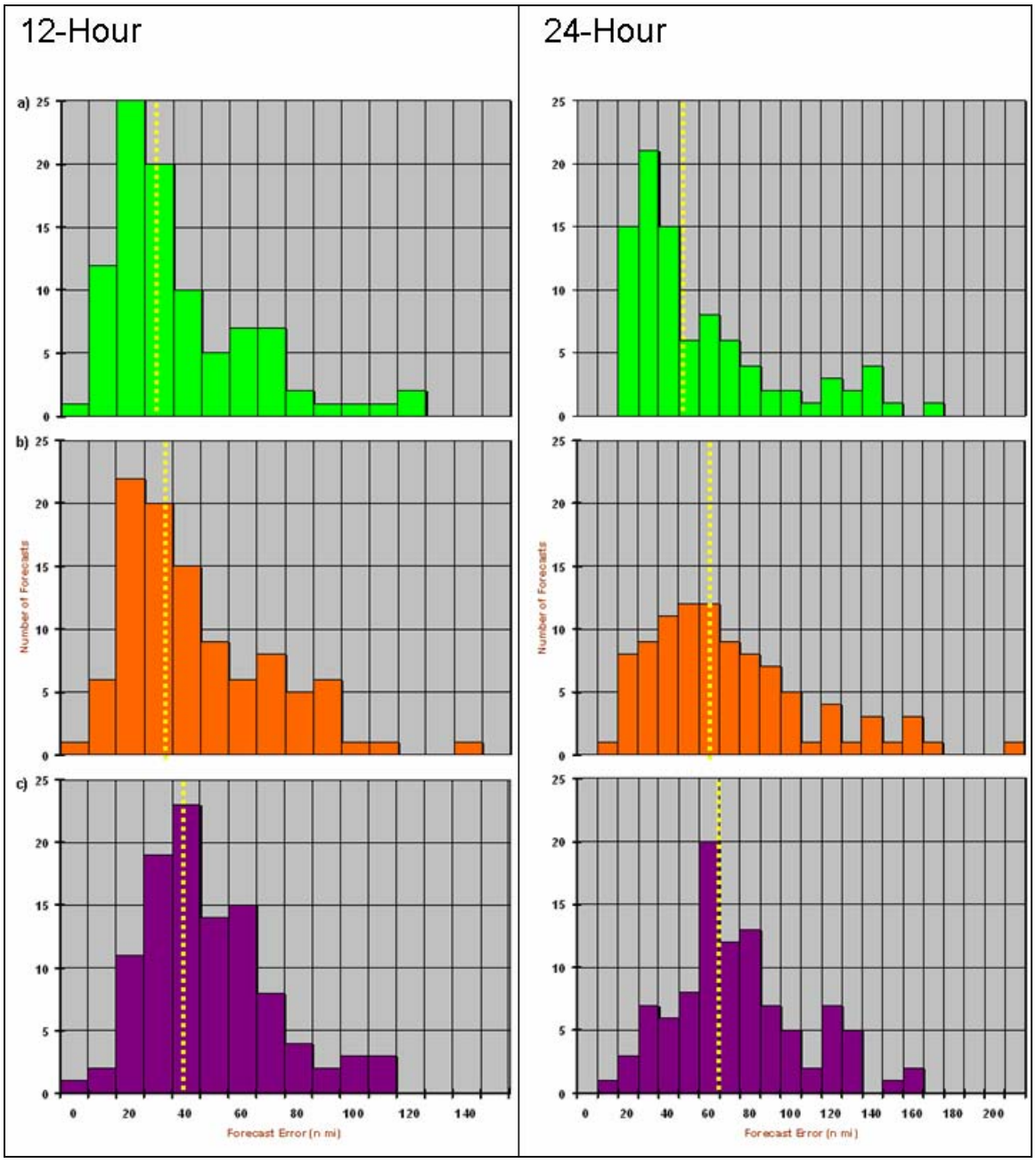

Figure 30. Histograms of 12- (left column) and 24-h (right column) OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

For the 72-, 96-, and 120-h forecasts, each has three failures out of the six comparison tests (Table 4). The error distributions become more random (Figures 32 and 33) and have smaller samples. While the 72- and 96-h histograms have enough samples

to create reasonable histograms, the 120-h interval does not (Figures 32 and 33) since only 18 samples per tercile are available for the 120-h tercile comparison.

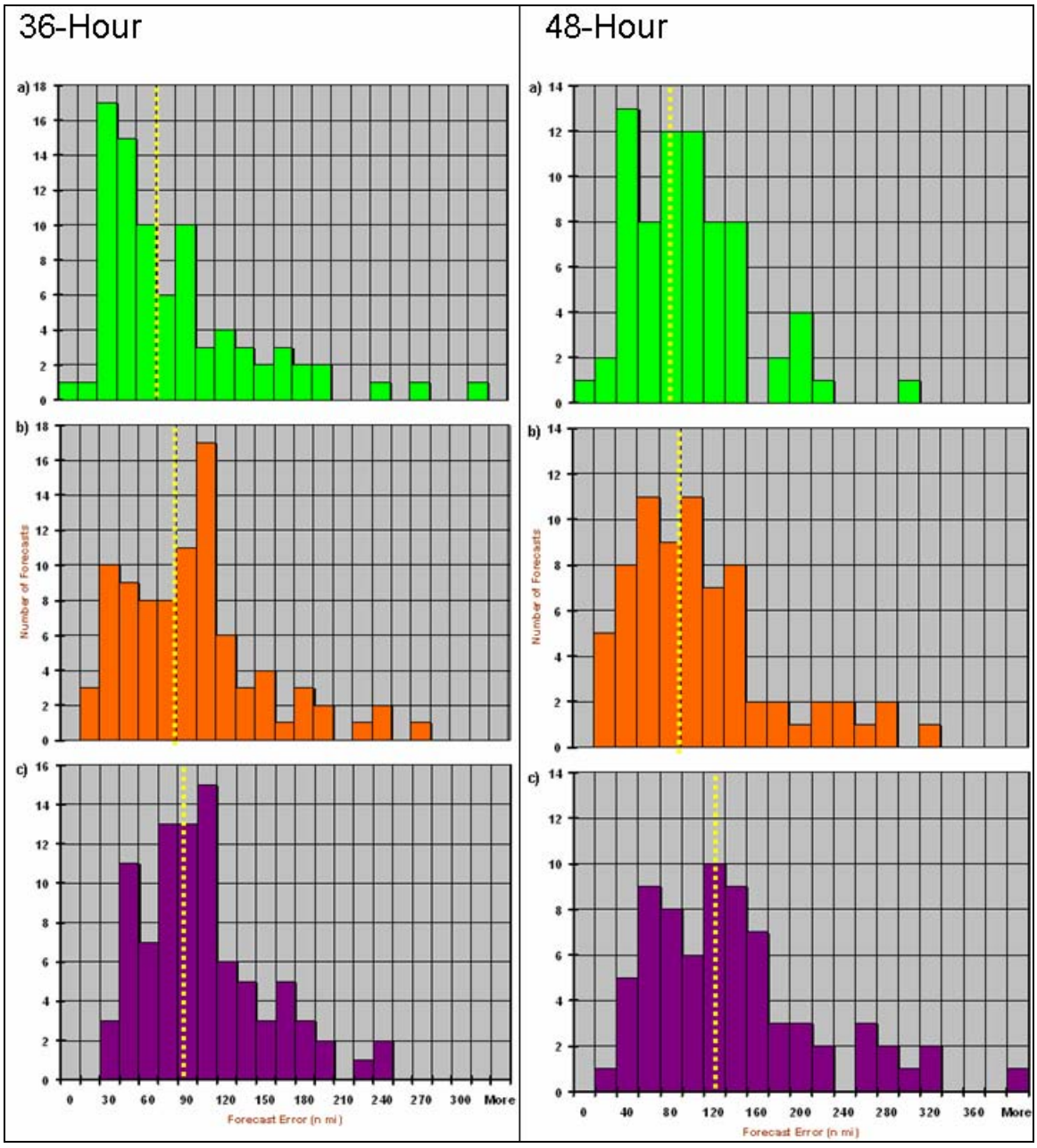

Figure 31. Histograms of 36- (left column) and 48-h (right column) OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

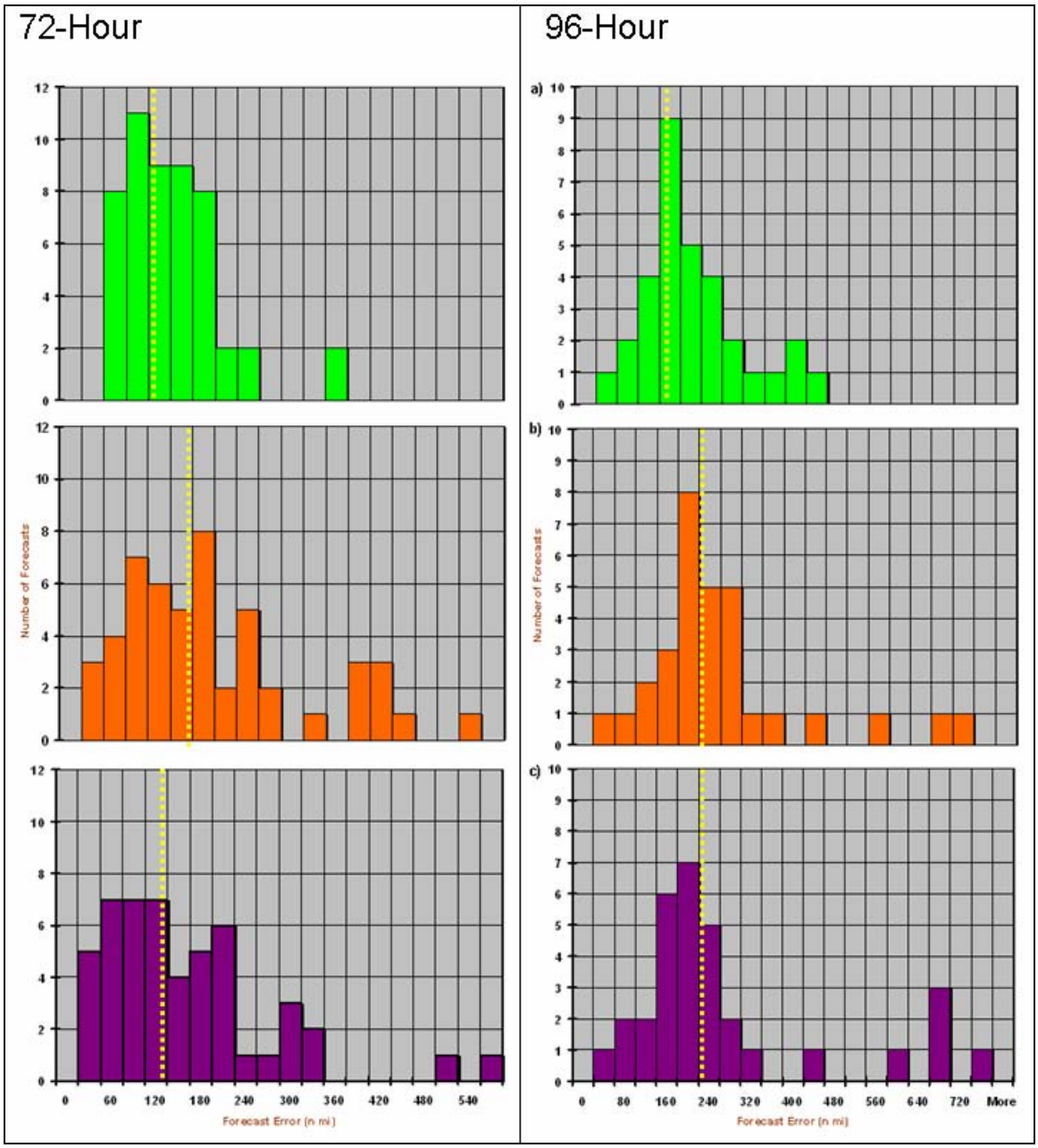

Figure 32. Histograms of 72- (left column) and 96-h (right column) OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

The skewness of the 72-h tercile distributions are to the right for the middle and low terciles and somewhat centered for the high tercile. While the 96-h terciles have two

highly skewed distributions in the middle and upper terciles, lower tercile has a smaller right-skewed distribution. The 120-h tercile distributions do not have enough samples to determine skewness.

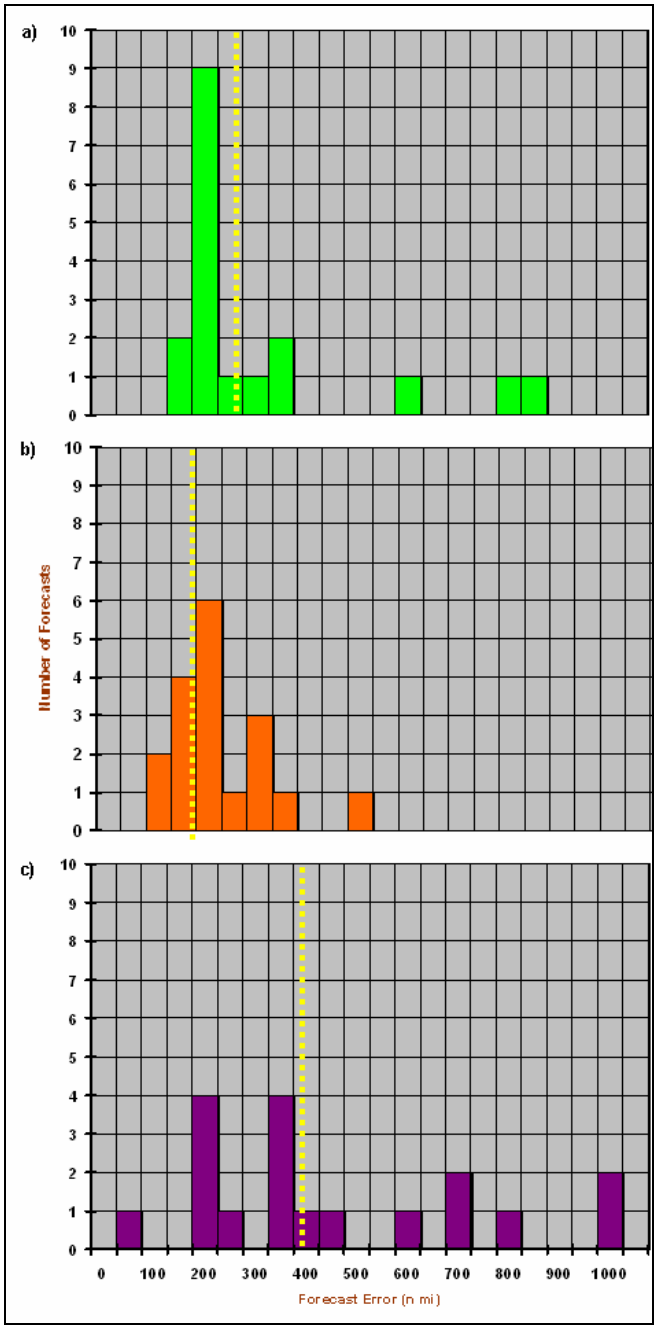

Figure 33. Histograms of 120-h OFCL total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

In the 72-, 96-, and 120-h forecast intervals, only the 96-h interval has increasing means and variances for decreasing forecast confidence (Table 4). The 72-h forecast interval low confidence tercile has a smaller mean and variance than the middle tercile, while in the 120-h forecast interval the middle tercile mean and variance are smaller than the other two.

2. Summary

The tests for differences in means resulted in only 8 out of 21 tercile comparisons having significantly different means. The tests for differences in variances were only slightly better with 10 of 21 comparisons having significantly different variances. The means and variances only increased steadily with decreasing forecast confidence for three of the seven forecast intervals. Additionally, only the 24-h forecast interval had the hypothesized progression of skewness from high to low forecast confidence.

The correlations between the GFS ensemble spread and OFCL total-track error were low for all forecast intervals with adjusted R^2 values between 0.01 and 0.04, except for 120 h which had a value of 0.12. In other words, for the 12- through 96-h forecast intervals, only a 1% to 4% variation in the OFCL total-track errors could be explained by the variations in the GFS ensemble spread.

With all these factors combined, it is clear that GFS ensemble spreads are not a good indicator of forecast confidence for the OFCL total-track forecast errors. Some of the negative results might be explained by the smaller samples with the longer forecast intervals. However, even with a larger number of samples in the shorter-range forecast intervals, there were few significant differences among the tercile distributions. Based on these results for the ensemble mean total-track errors, using the GFS ensemble spread to condition OFCL along- and cross- track errors was not examined.

F. GFS ENSEMBLE MEAN TOTAL-TRACK FORECAST ERRORS CONDITIONED ON GFS ENSEMBLE SPREADS

Since the GFS ensemble spread performed poorly as a measure of forecast confidence for the OFCL total-track forecast errors, tests were conducted to determine whether the GFS ensemble spread is even a good indicator of forecast confidence for GFS ensemble mean total-track errors. The GFS ensemble mean total-track errors were binned using GFS ensemble spread using the same method discussed in the previous

section. The results of the comparisons are summarized in Table 5, and the tercile comparison histograms are displayed in Figures 34 through 37.

Table 5. The tercile comparison table for the GFS ensemble mean total-track forecast errors conditioned on GFS ensemble spreads. The legend in the upper right portion of the table defines the color scheme and tercile definitions.

Forecast Interval	12-H Forecast			GFS Ensemble Mean Total-Track Forecast Errors Conditioned on GFS Ensemble Spread		
Tercile	Lower	Middle	Upper			
Samples	94	101	105	Legend		
Mean (n mi)	34.1	45.1	46.6	Statistically Different (at 0.05 alpha)	Within 1% C.L.of Pass/Fail	Statistically the same (at 0.05 alpha)
Standard Deviation (n mi)	24.6	40.9	25.2			
Total Distribution	Samples: 300 Mean: 42.2 SD: 31.7 R: 0.20 R^2: 0.04 P: 0			Lower(L) - Lower tercile of track forecast errors when forecast confidence was high		
Comparison	L vs. M	M vs. U	L vs. U	Middle(M) - Middle tercile of track forecast errors when forecast confidence was average		
Test for differences in means	t-Stat: 2.44 P: 0.02	t-Stat: 0.34 P: 0.74	t-Stat: 3.73 P: 0.00			
Test for differences in variances	F-Stat: 2.77 P: 0.00	F-Stat: 2.63 P: 0.00	F-Stat: 1.05 P: 0.40	Upper(U) - Upper tercile of track forecast errors when forecast confidence was low		

Forecast Interval	24-H Forecast			36-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	100	104	105	87	87	90
Mean (n mi)	54.7	57.3	65.6	68.4	77.6	88
Standard Deviation (n mi)	33.9	35.8	36.7	43.6	48.5	44.5
Total Distribution	Samples: 309 Mean: 59.3 SD: 35.7 R: 0.16 R^2: 0.02 P: 0			Samples: 264 Mean: 78.1 SD: 46.1 R: 0.15 R^2: 0.02 P: 0.01		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 0.52 P: 0.60	t-Stat: 1.65 P: 0.10	t-Stat: 2.19 P: 0.01	t-Stat: 1.31 P: 0.19	t-Stat: 1.49 P: 0.14	t-Stat: 2.96 P: 0.00
Test for differences in variances	F-Stat: 1.11 P: 0.30	F-Stat: 1.05 P: 0.40	F-Stat: 1.17 P: 0.21	F-Stat: 1.24 P: 0.16	F-Stat: 1.19 P: 0.21	F-Stat: 1.04 P: 0.43

Forecast Interval	48-H Forecast			72-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	74	69	75	50	51	52
Mean (n mi)	80.5	85.3	111	107	134	146
Standard Deviation (n mi)	51.9	53.5	62.5	53.4	78.6	90.6
Total Distribution	Samples: 218 Mean: 92.6 SD: 57.6 R: 0.22 R^2: 0.04 P: 0			Samples: 153 Mean: 129 SD: 77.3 R: 0.20 R^2: 0.01 P: 0.01		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 0.55 P: 0.59	t-Stat: 2.69 P: 0.01	t-Stat: 3.27 P: 0.00	t-Stat: 2.05 P: 0.04	t-Stat: 0.71 P: 0.48	t-Stat: 2.68 P: 0.00
Test for differences in variances	F-Stat: 1.06 P: 0.40	F-Stat: 1.37 P: 0.10	F-Stat: 1.45 P: 0.06	F-Stat: 2.17 P: 0.00	F-Stat: 1.33 P: 0.16	F-Stat: 2.88 P: 0.00

Forecast Interval	96-H Forecast			120-H Forecast		
Tercile	Lower	Middle	Upper	Lower	Middle	Upper
Samples	33	32	33	18	18	19
Mean (n mi)	158	203	248	180	245	308
Standard Deviation (n mi)	71.4	106	160	141	109	167
Total Distribution	Samples: 98 Mean: 203 SD: 123 R: 0.45 R^2: 0.20 P: 0			Samples: 55 Mean: 245 SD: 149 R: 0.41 R^2: 0.17 P: 0		
Comparison	L vs. M	M vs. U	L vs. U	L vs. M	M vs. U	L vs. U
Test for differences in means	t-Stat: 1.98 P: 0.05	t-Stat: 1.35 P: 0.18	t-Stat: 2.95 P: 0.01	t-Stat: 1.54 P: 0.13	t-Stat: 1.36 P: 0.18	t-Stat: 2.52 P: 0.02
Test for differences in variances	F-Stat: 2.21 P: 0.01	F-Stat: 2.63 P: 0.00	F-Stat: 5.02 P: 0.00	F-Stat: 1.68 P: 0.15	F-Stat: 2.34 P: 0.04	F-Stat: 1.40 P: 0.25

1. Analysis and Results

The GFS ensemble spreads provided a slightly better measure of forecast confidence for the ensemble mean total-track forecast errors than they did for the OFCL errors. The tercile histograms in Figures 34 through 37 indicate that for every forecast interval the total-track error means increase with decreasing forecast confidence.

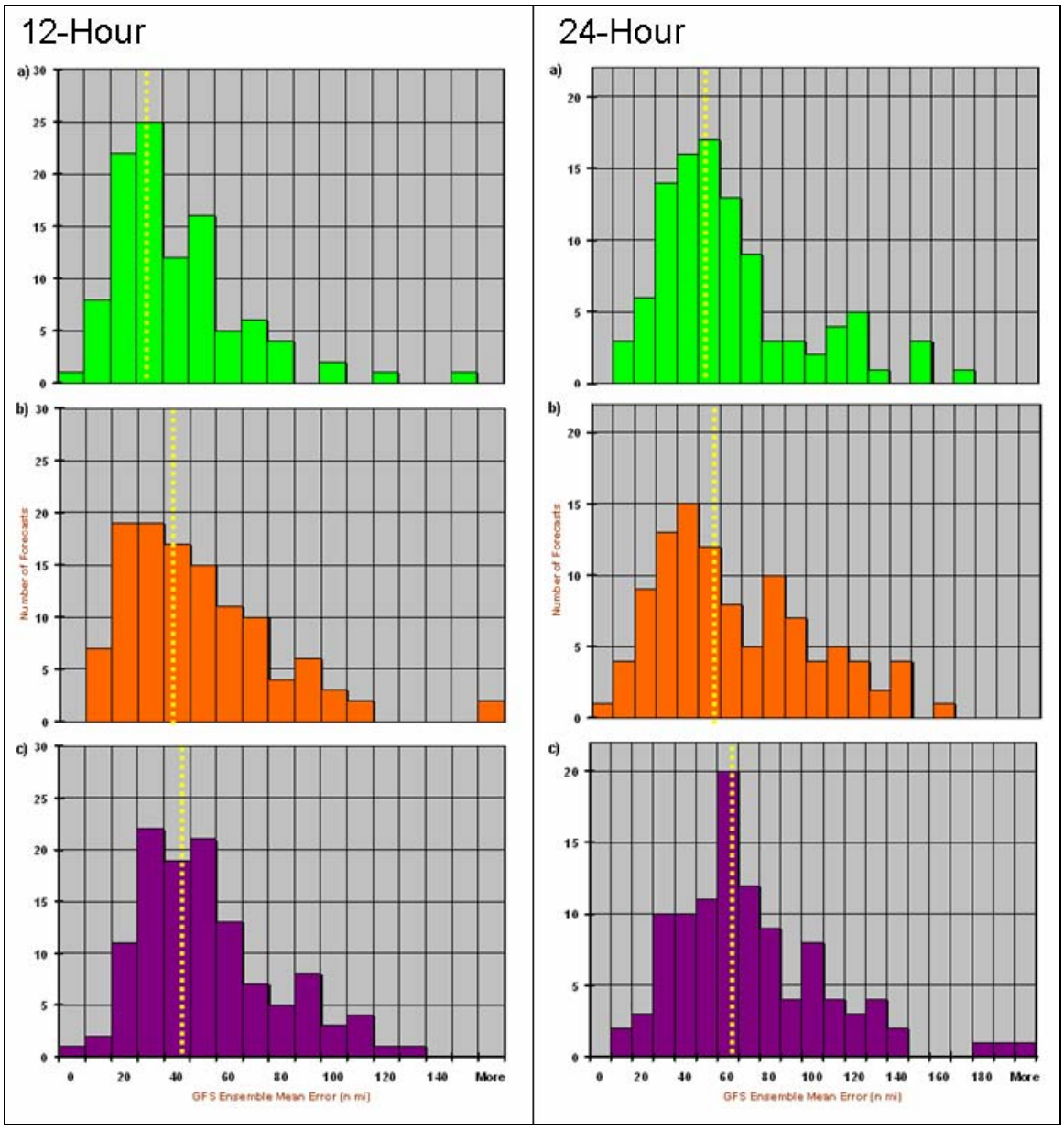

Figure 34. Histograms of 12- (left column) and 24-h (right column) GFS ensemble mean total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

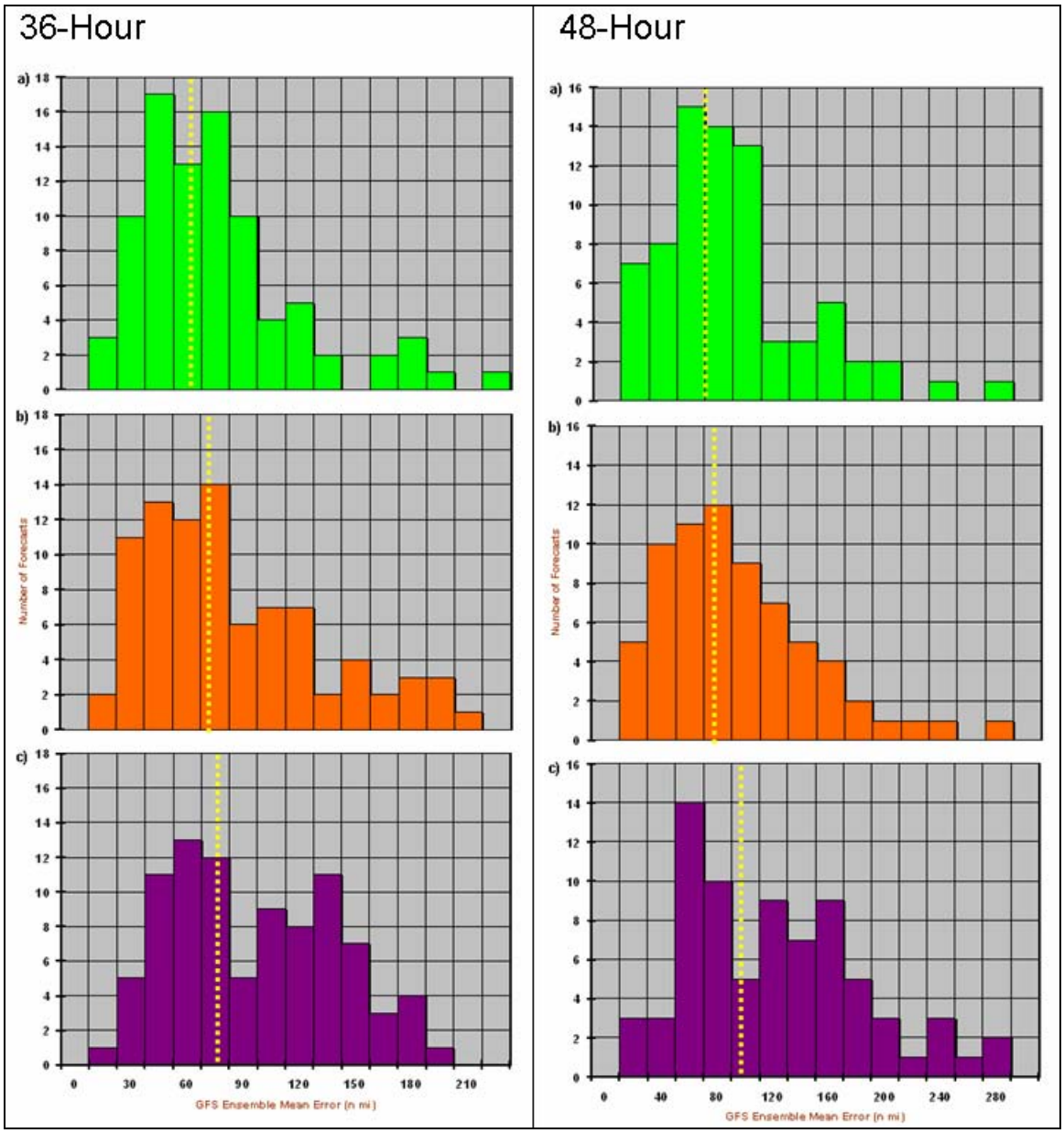

Figure 35. Histograms of 36- (left column) and 48-h (right column) GFS ensemble mean total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

However, the results of the statistical tests were not any better (Table 5) than the results from using the GFS ensemble spreads as a measure of forecast confidence to condition the OFCL total-track forecast errors (Table 4). The tests for differences in means resulted in only 11 of the 21 comparisons having significantly different means,

while the tests for differences in variances had only 8 out of 21 comparisons having significantly different variances. In addition, the variances do not increase steadily with decreasing forecast confidence in three of the seven ensemble mean track error distributions. As in the previous section, only the 24-h forecast interval had the hypothesized progression of skewness from high to low forecast confidence (Figure 34, 24-h panels a to c).

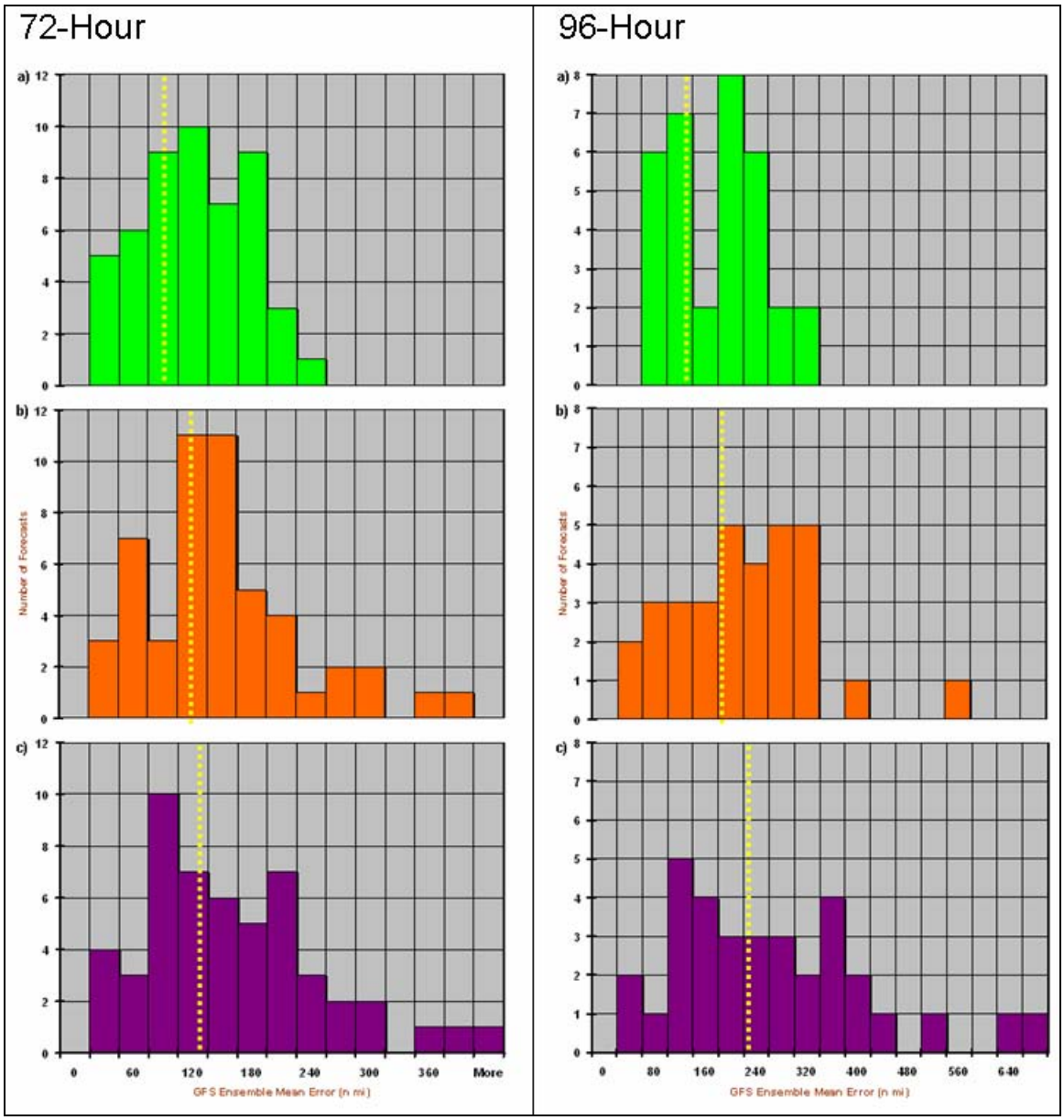

Figure 36. Histograms of 72- (left column) and 96-h (right column) GFS ensemble mean total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

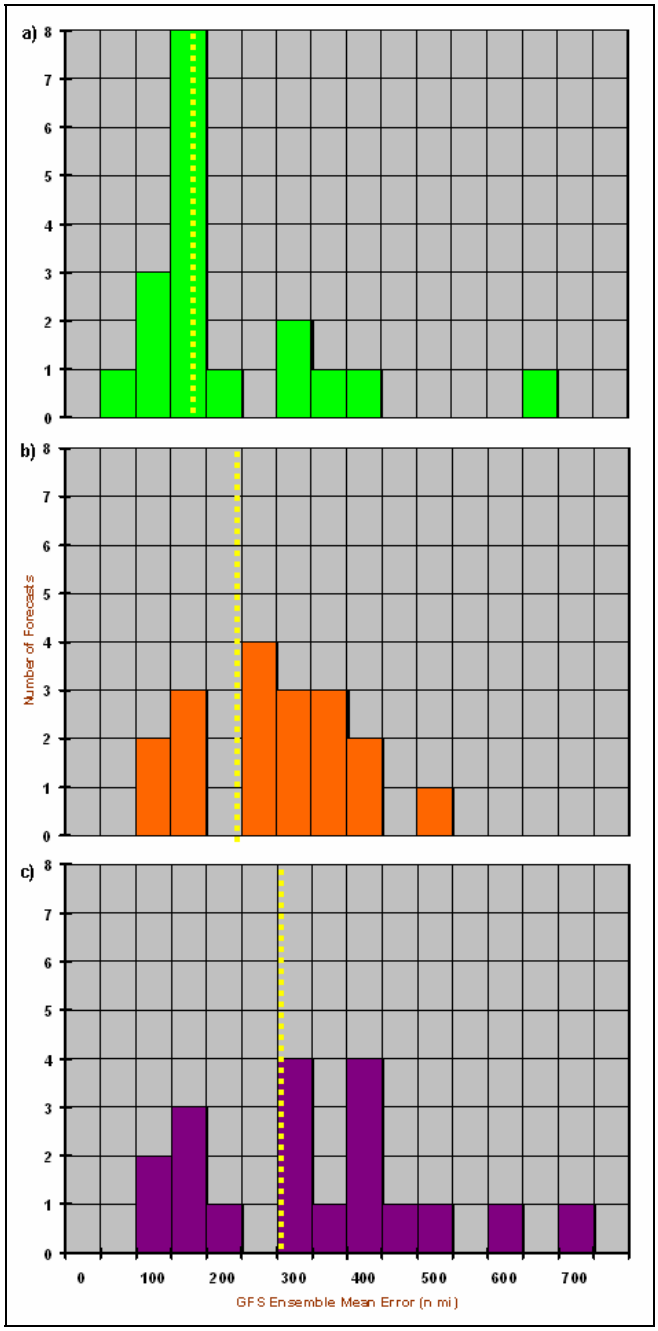

Figure 37. Histograms of 120-h GFS ensemble mean total-track forecast error distributions conditioned on a) high, b) average, and c) low forecast confidence. The means are represented by the dashed lines.

2. Summary

There was very little indication that GFS ensemble spread is a good measure of forecast confidence for the GFS ensemble mean total-track forecast errors. Rather, the conclusion is that using the spread of a consensus via GPCE values instead of a single model ensemble spread is more reliable in determining overall forecast confidence of TC track forecasts.

G. DIFFERENT INDICATORS OF FORECAST CONFIDENCE VS TRACK ERRORS FOR HURRICANE WILMA

To further illustrate the differences between using GPCE values rather than GFS ensemble spreads as a measure of forecast confidence, the track forecasts for Hurricane Wilma (October 2005) were examined. The 72-h forecast interval was chosen since it is near the middle of the seven forecast periods and is representative of the differences at other forecast intervals.

The GPCE values and OFCL total-track forecast errors (Figure 38), GFS ensemble spreads and OFCL total-track forecast errors (Figure 39), and the GFS ensemble spreads and the GFS ensemble mean total-track errors (Figure 40) were compared to illustrate how changes in the measures of forecast confidence varied with track errors. The highest correlation of 0.51 was between the GPCE values and the OFCL total-track forecast errors. The next highest correlation was 0.38 for the GFS ensemble spreads and the GFS ensemble mean total-track forecast errors. The lowest correlation of 0.25 was between the GFS ensemble spreads and the OFCL total-track forecast errors. These results are consistent with the earlier seasonal summaries, and illustrate again that the GPCE values provide a better measure of OFCL forecast confidence than the GFS ensemble spreads.

In the two comparisons involving OFCL total-track forecast errors for Hurricane Wilma in Figures 38 and 39, the forecast confidence indicators started at a higher value and were poorly correlated with the track errors for the first 18 to 24 hours, perhaps because the storm was in its early stages when the forecast models normally have less skill. Both confidence indicators seem to be out of phase with the track error.

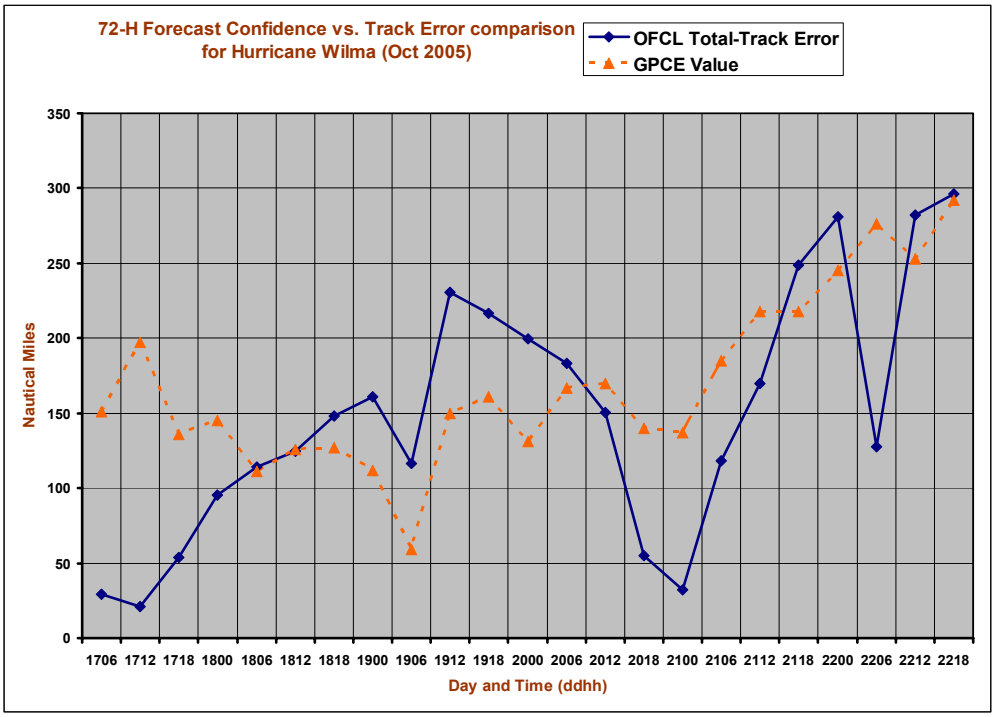

Figure 38. The 72-h OFCL total-track forecast errors and GPCE values (ordinate) at the times for each advisory (abscissa) issued for Hurricane Wilma (2005).

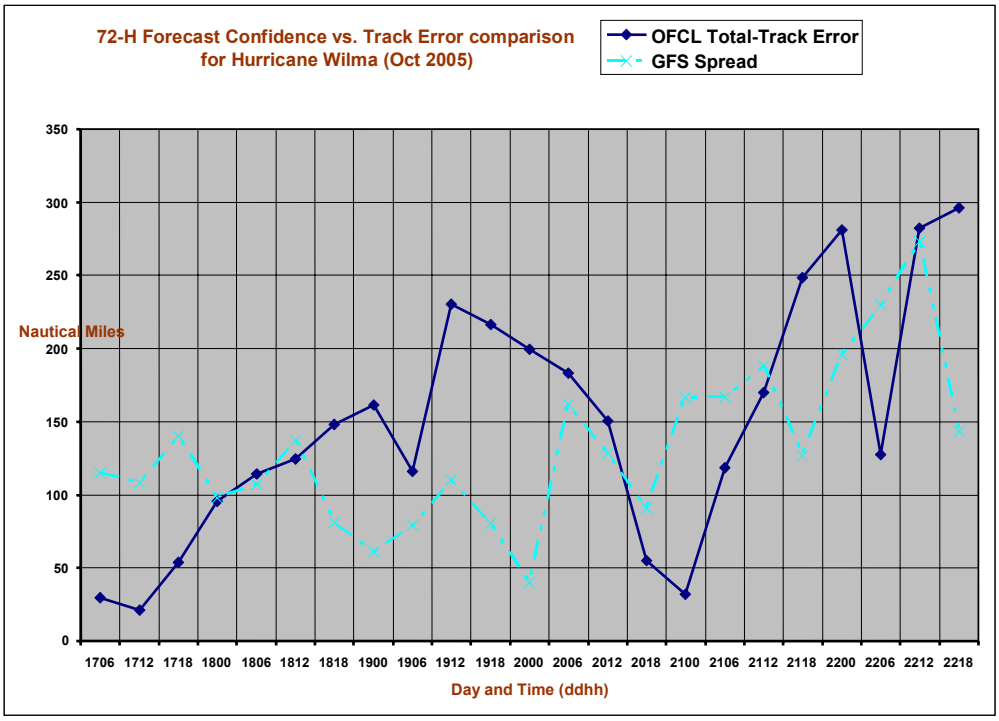

Figure 39. The 72-h OFCL total-track forecast errors and GFS ensemble spreads (ordinate) at the times for each advisory (abscissa) issued for Hurricane Wilma (2005).

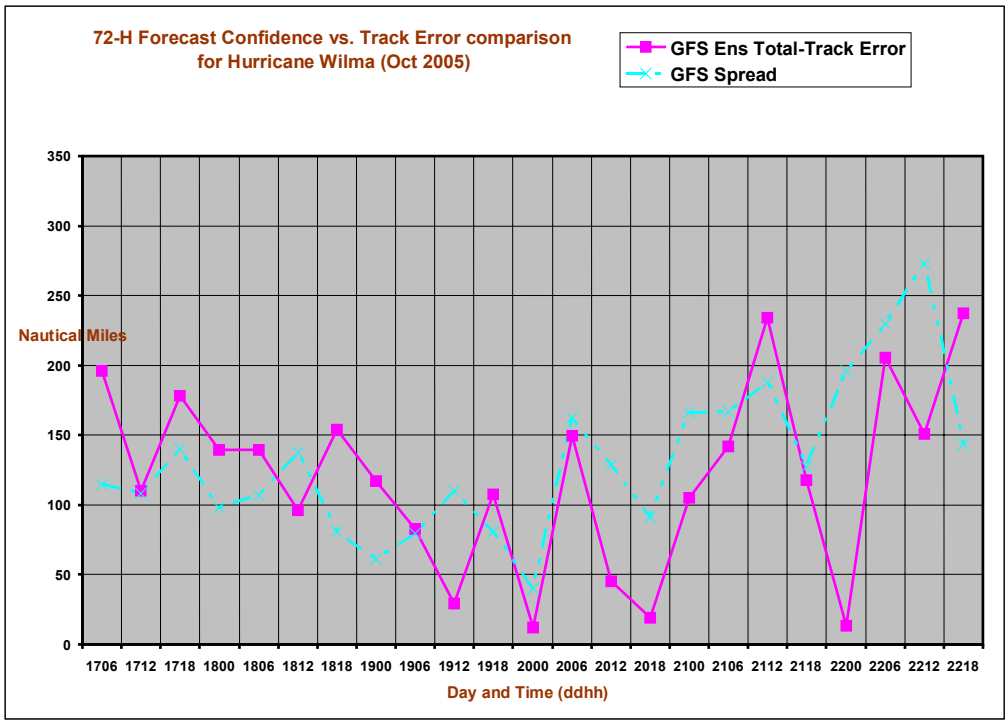

Figure 40. The 72-h GFS ensemble mean total-track forecast errors and GFS ensemble spreads (ordinate) at the times for each advisory (abscissa) issued for Hurricane Wilma (2005).

After the poor start, the GPCE values have a reasonably good correlation with the OFCL total-track forecast error (Figure 38). For example, a large decrease in track forecast error coincided with a similar decrease in GPCE value for the 0600 UTC 19 October forecast and the 0000 UTC forecast on 21 October. Also the GPCE values were a reasonable indicator of the increasing OFCL track forecast errors from the 0600 UTC 21 October forecast through the last forecast at 1800 UTC 22 October. However, large decrease in track error at 1200 UTC 22 October was not predicted by the GPCE value. However, such a large departure from the general trend in errors may simply be an excellent OFCL forecast that departed from the consensus track guidance.

Although the GFS ensemble spreads generally increase with OFCL track forecast errors (Figure 39), the correlation is not as good for the GPCE values. In some cases, the GFS ensemble spread is out of phase with the total-track forecast error. Some examples are the 0600 UTC 19 October forecast and the 0000 UTC 21 October forecast in which the OFCL error and GPCE value both decreased, but the GFS ensemble spread had

decreased 6 hours earlier. The GFS ensemble spreads also have more variability than the GPCE values. After 1200 UTC 19 October, the largest variability in the GPCE values relative to the general trend is 25 n mi on October 22. By contrast, the GFS ensemble spreads routinely vary by 75 n mi from one forecast period to the next, including a 90 n mi variation on 22 October.

The GFS ensemble mean track forecast errors versus GFS ensemble spreads often have large variability from one forecast period to the next throughout the period of Hurricane Wilma (Figure 40). At times, the GFS ensemble mean errors and spreads are in phase, and at other times they are not. This lack of correlation again illustrates that not only are the GFS ensemble spreads a poor indicator of OFCL track forecast confidence, but also are a poor indicator of its own ensemble mean forecast track error.

V. CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

This study was a first step in an investigation of whether a new version of the MC model for occurrences of tropical cyclone-induced wind would be improved by conditioning the track error distribution that is used to calculate probabilities to account for the track forecast difficulty. If it were possible to use different track error distributions for different levels of forecast confidence, the probability wind output may be more accurate. For example, a reduction in the massive costs of an overly cautious evacuation may be possible when the track forecast confidence is high, or even save lives by expanding the evacuation zone when the track forecast confidence is low.

In this first step, two predictors were tested as measures of track forecast confidence. One predictor was the GPCE value, which is calculated from the spread of the individual member tracks in the CONU model. The other predictor was the GFS ensemble model track spread, which is calculated from the average distance of the individual ensemble members from the ensemble mean. The GPCE values were used as predictors of the magnitude of the OFCL total-track forecast errors, cross-track forecast errors, and along-track forecast errors. Similarly, the GFS ensemble spreads were used as an indicator of forecast confidence in the OFCL total-track forecast error and in the GFS ensemble mean total-track forecast error.

Statistical techniques were used to determine if the means and the variances of these track forecast errors were significantly different when the terciles of GPCE values and GFS ensemble spreads were used as predictors of forecast confidence. That is, would the distributions of track forecast errors be significantly different in three distributions of track forecast errors?

1. Official Total-, Along-, and Cross-Track Forecast Errors Conditioned on GPCE Values

The OFCL total-track forecast errors were examined first to provide a possible basis for investigating the along- and cross-track forecast errors. The tests for differences in means resulted in 19 of the 21 tercile comparisons having significantly different means. Similarly, the results of the tests for differences in variances showed that 18 of

the 21 tercile comparisons had significantly different variances (Table 1). Additionally, the OFCL track error means and variances increased as forecast confidence decreased for all forecast intervals. For all but the 12-h forecast interval, the tercile distributions followed the hypothesized progression of increased skewness from high forecast confidence to low. The correlations of the GPCE values and track errors at the various forecast intervals ranged from 0.29 to 0.41, which were consistent with expectations. These results indicate that GPCE values are a good indicator of forecast confidence. This test with total-track errors warranted a closer look at the along- and cross-track forecast error components of the OFCL forecast errors from which the MC model draws its historic track errors.

The OFCL along-track forecast errors were tested to see if they produced similar results to the OFCL total-track forecast errors. The variance tests resulted in 18 of the 21 tercile comparisons having significantly different variances (Table 2). Two of the failures were consistent with the total-track test failures for the 72-h middle- and upper-terciles and 120-h lower- and middle-terciles comparisons. Although the failure of the 36-h middle- and upper-terciles along-track error comparison did not have a corresponding failure with total-track errors, that comparison did have a P-value of 0.03. Conversely, the 24-h along-track error lower and middle terciles comparison passed while the corresponding total-track error terciles comparison failed. As for the OFCL total-track forecast errors, the along-track forecast errors increased in variance as forecast confidence decreased for all forecast periods. Taking these factors into account along with the results of the total-track error comparisons, it is clear that OFCL along-track forecast errors can be successfully stratified by forecast confidence based on the GPCE values.

Next, the OFCL cross-track forecast errors were tested to see if they were consistent with the results of the total- and along-cross error tercile comparisons. The differences in variances test resulted in 16 of the 21 comparisons having significantly different variances (Table 3). The only test failures for the cross-track forecast errors were also those for the along- and total-track errors, e.g., the 120-h lower- and middle-terciles comparison (Table 1 and 2). The 36-h middle- and upper-tercile comparison P-value of 0.5 is consistent with the along-track error test failure for the same comparison.

The 96-h cross-track error test failure for the middle- and upper-tercile comparison is consistent with the same total-track error comparison, which resulted in a P-value of 0.05. Finally, the 12- and 48-h terciles comparison test failure was unique compared to the cross-track errors. Cross-track errors for the OFCL are usually smaller in magnitude than the along-track errors. Therefore, the range of the cross-track distributions was smaller than the along-track distributions, which means there is less room for variability. This difference in variability resulted in some of the tests for differences in variances for the along-track errors having a slightly higher P-value than the cross-track results. With the exception of the 12-h forecast interval, all other forecast intervals had increasing variance with decreasing forecast confidence, which indicates that for the majority of the forecast intervals, that as forecast confidence lowers, forecast track error becomes less predictable. As for the along-track errors, if the MC model drew from three cross-track error distributions based on forecast confidence, the area covered by each probability interval should increase with decreasing forecast confidence.

These tests using the GPCE values as a measure of forecast confidence to condition OFCL along- and cross-track forecast errors indicate that using this method will successfully stratify the errors into significantly different distributions. The MC model will most likely benefit from adopting this approach instead of drawing from just one static distribution.

2. Testing the Effectiveness of GFS Ensemble Spread as an Indicator of Forecast Confidence

Using GFS ensemble spreads as a measure of forecast confidence to condition OFCL total-track forecast errors was also tested to see if the results were any better then using GPCE values. These GFS ensemble-based tests for differences in means resulted in only 8 of 21 comparisons having significantly different means, while the tests for differences in variances had only 10 of 21 comparisons having significantly different variances. The means and variances only increased steadily with decreasing forecast confidence for three of the seven forecast intervals. Additionally, only the 24-h forecast interval had the hypothesized progression of skewness from high to low forecast confidence. Furthermore, the correlations between the GFS ensemble spread and OFCL total-track error were small for all forecast intervals with adjusted R^2 values between 0.01

and 0.04, except for the 120-h interval that had a 0.12. These tests indicate that GFS ensemble spreads are not a good indicator of forecast confidence for OFCL total-track forecast errors. Therefore, using the GFS ensemble spreads to condition OFCL along- and cross-track errors was not examined.

It was also investigated whether the GFS ensemble spread provided an accurate indicator of forecast confidence for the ensemble track-error mean. The statistical tests indicated no significant improvement. The tests for differences in means resulted in only 11 of 21 comparisons having significantly different means, while the tests for differences in variances had only 8 of 21 comparisons having significantly different variances. In addition, the ensemble track error variances did not increase steadily with decreasing forecast confidence in three of the seven distributions. The only consistent positive result was that the means for the tercile distributions increased with decreasing forecast confidence for every forecast interval.

It is concluded from these two comparisons using the GFS ensemble spread as a measure of forecast confidence that the MC model would not benefit from using the GFS ensemble spread to stratify OFCL along- and cross-track errors. Rather, it is concluded that using the spread of a consensus instead of a single model ensemble spread is more reliable method for specifying overall forecast confidence of TC track forecasts.

3. Summary

The key result of this thesis is that changing the MC model of the new NHC probabilistic product to draw from one of three historic OFCL along- and cross-track forecast error distributions conditioned on forecast confidence derived from the GPCE values will most likely yield improved results. However, using GFS ensemble spread as a measure of forecast confidence will not improve the model, and may actually degrade it.

Such a change in the new probabilistic model is expected to improve the accuracy of its probabilistic wind output. This change could lead to a reduction of the massive costs of overly cautious evacuations when forecast confidence is high, or even save lives

by expanding the evacuation zone when forecast confidence is low. In the end, the changes will result in improved TC forecasting support for both the military and civilian sectors.

B. RECOMMENDATIONS

The next step would be to change the MC model code to draw from three along-track and cross-track error distributions based on forecast confidence (GPCE value) instead of just one error distribution for all track forecast situations. The model should then be tested to see if the probabilistic wind distribution accuracy is significantly improved. If so, a permanent change in the model should be made.

Future research should concentrate on other factors that may influence historic track error distributions. Some of those factors may be the time of year, steering flow characteristics, storm intensity, weather regime of the eastern U.S., and TC origin.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Elsberry, R. L., and L. E. Carr, 2000: Consensus of dynamical tropical cyclone track forecasts—Errors versus spread. *Mon. Wea. Rev.*, **128**, 4131–4138.

Grimit, E. P., and C. F. Mass, cited 2006: Measuring the ensemble spread-error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.*, accepted. [Available online at http://www.atmos.washington.edu/~ens/pdf/spread_errorI.version2.0.pdf.]

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*., **128**, 1187–1193.

———, 2005: Quantifying tropical cyclone track forecast uncertainty and improving extended-range tropical cyclone track forecasts using an ensemble of dynamic models, Semi-annual Report for JHT Project, 93 pp. [Available online at http://www.nhc.noaa.gov/jht/2003-2005reports/jht_final_goerss.pdf.]

———, J. S., 2006: Prediction of consensus tropical cyclone track forecast error for hurricanes Katrina, Rita, and Wilma. *Extended Abstracts, 27th Conf. on Hurricanes and Tropical Meteorology*, Monterey, CA, Amer. Meteor. Soc., 11A.1.

Gross, J. M., M. DeMaria, J. A. Knaff, and C. R. Sampson, 2004: A new method for determining tropical cyclone wind forecast probabilities. *Extended Abstracts, 26th Conf. on Hurricanes and Tropical Meteorology*, Miami, FL., Amer. Meteor. Soc., 425-426.

Knaff, J. A., and M. DeMaria, 2005: Improvements in deterministic and probabilistic tropical cyclone surface wind predictions. JHT Final Report, 7 pp. [Available from Cooperative Institute for Research in the Atmosphere, Colorado State University, West Laporte Avenue, Fort Collins, CO 80523-1375].

———, and M. DeMaria, 2006: Continued development of tropical cyclone wind probability products. Proposal to NOAA Joint Hurricane Testbed (JHT) Opportunities for Transfer of Research and Technology into Tropical Cyclone Analysis and Forecast Operations, 8 pp. [Available from Cooperative Institute for Research in the Atmosphere, Colorado State University, West Laporte Avenue, Fort Collins, CO 80523-1375].

Leslie, L. M., and K. Fraedrich, 1990: Reduction of tropical cyclone position errors using an optimal combination of independent forecasts. *Wea. Forecasting*, **5**, 158-161.

Neumann, C. J., and J. M. Pelissier, 1981: An analysis of Atlantic tropical cyclone forecast errors, 1970-1979. *Mon. Wea. Rev*., **109**, 1248-1266.

NHC, cited 2006: Tropical Weather Summary. [Available online at http://www.nhc.noaa.gov/archive/2005/tws/MIATWSAT_nov.shtml.]

NOAA, cited 2006: Hurricane evacuation zone maps. [Available online at http://www.dem.dcc.state.nc.us/hurricane/HurricaneEvacuationRoutes.pdf.]

NPMOC, cited 2006: Summary of forecast verification. [Available online at http://www.npmoc.navy.mil/jtwc/atcr/1998atcr/ch5/chap5_page1.html.]

Thompson, P.D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev*., **105**, 228-229.

Trochim, W. M., cited 2006: Research methods knowledge base. [Available online at http://www.socialresearchmethods.net/kb/.]

Wadsworth, cited 2006: Workshop: Independent Verses Repeated t Tests. [Available online at http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/ttest_betwn/ttest_betwn_02.html.]

Whitehead, J. C., 2000: One million dollars a mile? The opportunity costs of hurricane evacuation. [Available online at http://ideas.repec.org/p/wop/eacaec/0005.html.]

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Science Second*. Elsevier Academic Press, 627 pp.

Wilson, J., cited 2006: Five names retired from the 2005 Season. [Available online at http://www.weather.com/newscenter/tropical/?from=wxcenter_news.]

Winters, K. A., 2006: Providing tropical cyclone weather support to space launch operations. *Extended Abstracts, 27th Conf. on Hurricanes and Tropical Meteorology*, Monterey, CA, Amer. Meteor. Soc., 9A.5.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Air Force Weather Technical Library
   Asheville, North Carolina

4. Professor Patrick Harr
   Naval Postgraduate School
   Monterey, California

5. Professor Russell Elsberry
   Naval Postgraduate School
   Monterey, California

6. Dr. Mark DeMaria
   National Environmental Satellite, Data, and Information Service
   Fort Collins. California

7. Dr. James Goerss
   Navy Research Laboratories
   Monterey, California

8. Mr. Buck Sampson
   Navy Research Laboratories
   Monterey, California

9. Mr. James Franklin
   National Hurricane Center
   Miami, Florida

10. Mr. William Roeder
    45 Weather Squadron
    Patrick Air Force Base, Florida

11. Captain Matthew Hauke
    21 Operational Weather Squadron
    Sembach, Germany